



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN  
PHILOSOPHY, SCIENCE, COGNITION, AND SEMIOTICS (PSCS)

Ciclo 37

**Settore Concorsuale:** 11/C2 - LOGICA, STORIA E FILOSOFIA DELLA SCIENZA.

**Settore Scientifico Disciplinare:** M-FIL/02 - LOGICA E FILOSOFIA DELLA SCIENZA

TU, ROBOT. VERSO NUOVI MODELLI DI INTERAZIONE TRA UMANI E ROBOT

**Presentata da:** Martina Bacaro

**Coordinatore Dottorato**

Claudio Paolucci

**Supervisore**

Francesco Bianchini

**Co-supervisore**

Claudio Paolucci

Esame finale anno 2025



## Abstract

Il rapporto tra umano e macchina è, ad oggi, una delle principali modalità attraverso cui viene portata avanti l'indagine sulla cognizione umana. Fin dal momento in cui è stato possibile costruire dispositivi artificiali che riuscissero a realizzare compiti cognitivi, alle macchine viene riservato il ruolo di testare, attraverso processi di modellizzazione e implementazione, le ipotesi sul funzionamento dei processi che definiscono l'intelligenza umana. La robotica, come campo disciplinare in cui si costruiscono artefatti che esibiscono livelli anche minimi di autonomia, costituisce una delle frontiere più avanzate di questa indagine, ultimo anello di una catena di realizzazioni che, negli ultimi settanta anni, è stata guidata dalla metafora della mente come computer. In questo contesto, sotto l'influenza delle teorie computazionali e funzionaliste della cognizione, il robot ha assunto il ruolo dell'omologo dell'essere umano, agente individuale la cui mente è pensata essere confinata all'interno del corpo o del cervello dell'individuo. Anche con l'avvento della robotica sociale, disciplina con lo scopo di dotare il robot di capacità di interazione con gli esseri umani, le architetture cognitive robotiche sono state costruite principalmente pensando all'individualità dell'essere umano, più che alla sua natura sociale. Il presente lavoro tenterà di avviare un'indagine inerente a modalità alternative di impiegare i robot per comprendere i fenomeni particolari della cognizione umana, mostrando come sia possibile non solo utilizzarli per testare ipotesi riguardo ai funzionamenti interni del nostro pensiero, ma anche le nostre modalità di interagire con gli altri; non solo macchine passive attraverso cui testare teorie individualiste, ma congegni che permettono la comprensione delle dinamiche interazionali. L'obiettivo principale di questa tesi consiste nel mostrare come approcci alternativi a quelli che hanno condotto alla realizzazione del robot sotto la guida della metafora computazionale possano ampliare lo sguardo sulla mente e instradare un'indagine in cui i robot ricoprono il ruolo di interlocutori in un dialogo. L'approccio adottato coincide con quello radicalmente *embodied* e situato delle teorie enattiviste della cognizione. Il campo di indagine in cui questo lavoro si situa è quello degli studi sull'interazione tra umani e robot, a sua volta composto di diverse anime che restituiscono un dominio intrinsecamente interdisciplinare. Attraverso il contributo congiunto di filosofia e scienze cognitive, si tenterà di sviluppare l'indagine sulla HRI prendendo in considerazione fenomeni al centro del campo disciplinare come la cognizione sociale, l'antropomorfismo, l'*embodiment*, l'attribuzione dell'intenzionalità e l'*Uncanny Valley Effect*. La tesi arriverà a mostrare come, spostando la prospettiva dai modelli della sostituzione, dell'individualità e della replicazione a quelli della collaborazione, del dialogo e dell'alterità sia possibile fornire interpretazioni alternative dei fenomeni in oggetto e riconfigurare l'indagine sotto l'insegna della partecipazione, che porta a vedere i robot non solamente come repliche dell'umano, ma come possibili partner nell'indagine sui processi cognitivi.

La tesi è strutturata come segue. Il *Capitolo I – Macchine come me. L'ambizione dell'umano come conoscenza attraverso la macchina* ricostruisce i presupposti epistemologici che hanno consentito alla macchina di assumere il ruolo di strumento di conoscenza privilegiato nell'indagine dell'essere umano. Il *Capitolo II – Dalle scienze cognitive alla robotica: andata e ritorno* fornisce una ricostruzione delle mutue contaminazioni tra i domini, analizzando i modi in cui ipotesi teoriche e implementazioni pratiche hanno modellato l'indagine sui processi cognitivi e sulle realizzazioni artificiali fino alla costruzione dei robot sociali. Il *Capitolo III – Galatea moderna. Robot sociali tra modellizzazione e mimesi del vivente* restituisce un'analisi dello stato dell'arte degli studi sulla HRI, con particolare attenzione ai modelli di cognizione sociale adottati, e fornisce un'analisi del robot come artefatto mimetico, sintetico e, al contempo, sincretico: si individuerà nell'ultimo aspetto la

cifra distintiva del robot sociale come macchina *sui generis*. Il *Capitolo IV – Umani, troppo umani. Il ruolo dell'intenzionalità nell'interazione tra umani e robot* prende in considerazione il fenomeno dell'attribuzione di intenzionalità ad artefatti robotici somiglianti all'umano, ne ricostruisce il dibattito interno agli HRI e fornisce un'analisi del robot come artefatto sincretico basata sulle teorie incarnate e situate della cognizione: riallacciando il legame tra cognizione sociale, agentività e comportamento, l'intenzionalità del robot verrà riletta sotto il rispetto del sincretismo presentato nel capitolo precedente. Il *Capitolo V – Il gioco dell'interazione. Rotture, abiti e aspettative nella HRI* elabora una variante del gioco dell'imitazione basata sulla reciprocità dell'interazione e sulla centralità dei *breakdown*: la teoria viene testata sull'*Uncanny Valley Effect* come fenomeno limite dell'esperienza umana di fronte al robot, concludendo con l'apertura di linee di ricerca nel campo dell'epistemologia e dell'apprendimento robotico, rese possibili dalla riconfigurazione dell'indagine sull'interazione tra umani e robot sotto il rispetto della corporeità, della socialità e della dipendenza dall'ambiente dei processi cognitivi umani.



*Milord, devo avvertirvi che stiamo per lasciare insieme i regni (inesplicati, certo, ma fin troppo battuti, no?) della vita normale, della Vita propriamente detta, per penetrare in un mondo di fenomeni insoliti e impressionanti. Vi darò la chiave del loro funzionamento. Ma già comincio dichiarandomi incapace, come il resto dell'Umanità (per il momento, almeno, e, temo, per sempre) di spiegare la natura di ciò che ne fa muovere i meccanismi.*

Auguste de Villiers de l'Isle-Adam,  
*Eva futura* (1886: 105)



## Indice

|                                                                                                                            |           |
|----------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Introduzione</b>                                                                                                        | <b>1</b>  |
| <b>Parte I - Storia ed epistemologia della robotica. Sullo sviluppo delle tecniche di mimesi e riproduzione dell'umano</b> | <b>15</b> |
| <b>Capitolo I – Macchine come me. L'ambizione dell'umano come conoscenza attraverso la macchina</b>                        | <b>16</b> |
| <b>1.1 L'ambizione dell'umano</b>                                                                                          | <b>18</b> |
| 1.1.1 Il mito della replicabilità                                                                                          | 22        |
| 1.1.2 Statue, miti e teste parlanti                                                                                        | 24        |
| <b>1.2 Il concetto di macchina</b>                                                                                         | <b>33</b> |
| 1.2.1 Macchine e automi                                                                                                    | 33        |
| 1.2.2 Macchine elettriche                                                                                                  | 44        |
| 1.2.3 Macchine cibernetiche                                                                                                | 47        |
| 1.2.4 Macchine computazionali                                                                                              | 56        |
| <b>1.3 Macchine: modelli per la mente</b>                                                                                  | <b>63</b> |
| <b>Capitolo II – Dalla robotica alle scienze cognitive: andata e ritorno</b>                                               | <b>67</b> |
| <b>2.1 Viaggio di andata: dalle scienze cognitive alla robotica</b>                                                        | <b>67</b> |
| 2.1.1 GOF AI: la ricetta per costruire robot                                                                               | 67        |
| 2.1.1.1 <i>L'ipotesi del Sistema Fisico di Simboli</i>                                                                     | 70        |
| 2.1.1.2 <i>Ricerca euristica</i>                                                                                           | 72        |
| 2.1.1.3 <i>Logica del razionale: SHRDLU</i>                                                                                | 74        |
| 2.1.2 Archeologia di SHAKEY                                                                                                | 78        |
| 2.1.2.1 <i>Il mondo di SHAKEY</i>                                                                                          | 79        |
| 2.1.2.2 <i>Livelli gerarchici di organizzazione: STRIPS</i>                                                                | 83        |
| <b>2.2 Realizzare un'idea: teoria e pratica della robotica</b>                                                             | <b>86</b> |
| <b>2.3 Viaggio di ritorno: dalla robotica alle scienze cognitive</b>                                                       | <b>91</b> |
| 2.3.1 Contra GOF AI                                                                                                        | 91        |
| 2.3.1.1 <i>Non farti ingannare dalle metafore: Joseph Weizenbaum</i>                                                       | 94        |
| 2.3.1.2 <i>Dalla logica alla biologia: Heinz von Foerster e Terry Winograd</i>                                             | 99        |
| 2.3.1.3 <i>Il corpo e l'esperienza: Hubert Dreyfus</i>                                                                     | 110       |
| 2.3.2 Nouvelle AI                                                                                                          | 118       |
| 2.3.2.1 <i>Behavior-based robotics</i>                                                                                     | 121       |

|         |                                                                                                             |            |
|---------|-------------------------------------------------------------------------------------------------------------|------------|
| 2.3.2.2 | <i>Cognitive Social Robotics</i>                                                                            | 128        |
| 2.3.3   | Modelli del mondo e modelli di mente                                                                        | 133        |
| 2.4     | <b>Nuove strade da intraprendere: filosofia della complessità e <i>Human-Robot Interaction</i></b>          | <b>139</b> |
|         | <b>Parte II - <i>Ti estì</i>. Il robot agli occhi dell'osservatore</b>                                      | <b>145</b> |
|         | <b>Capitolo III – Galatea Moderna. Robot sociali tra modellizzazione e imitazione del vivente</b>           | <b>146</b> |
| 3.1     | <b>Epistemologia di una nuova interazione</b>                                                               | <b>146</b> |
| 3.2     | <b>Human-Robot Interaction: questioni e prospettive</b>                                                     | <b>150</b> |
| 3.2.1   | Che cosa è un robot sociale?                                                                                | 152        |
| 3.2.2   | Come si dà l'interazione tra robot ed esseri umani?                                                         | 156        |
| 3.2.3   | Un dominio eterogeneo                                                                                       | 158        |
| 3.2.4   | Minimo comun denominatore: l'indagine della socialità                                                       | 163        |
| 3.3     | <b>Scienze cognitive, cognizione sociale e HRI</b>                                                          | <b>165</b> |
| 3.3.1   | Le persone adottano la ToM per spiegare e prevedere il comportamento del robot?                             | 168        |
| 3.3.2   | Come si può dotare il robot di una ToM?                                                                     | 170        |
| 3.3.3   | Not the only game in town                                                                                   | 171        |
| 3.3.4   | Enattivismo e cognizione sociale: dall'inferenza all'interazione                                            | 178        |
| 3.3.4.1 | <i>Alcune conseguenze per la HRI</i>                                                                        | 185        |
| 3.4     | <b>Interazione e nuovi agenti: per una HRI basata sulla HRI</b>                                             | <b>187</b> |
| 3.4.1   | Il robot come artefatto mimetico: imitazione estetica e antropomorfismo                                     | 191        |
| 3.4.1.1 | <i>Non è quel che sembra: l'antropomorfismo come errore cognitivo</i>                                       | 192        |
| 3.4.1.2 | <i>Robot come causa ed effetto della mimesi: l'antropomorfismo come fenomeno emergente dall'interazione</i> | 200        |
| 3.4.2   | Il robot come artefatto sintetico: la questione del comportamento e il ruolo dell'embodiment                | 208        |
| 3.4.2.1 | <i>Il Perceptual Belief Problem</i>                                                                         | 215        |
| 3.5     | <b>Il robot come artefatto sincretico: dal modello unidirezionale all'anello interazionale</b>              | <b>221</b> |
|         | <b>Capitolo IV – Umani, troppo umani. Il ruolo dell'intenzionalità nell'interazione tra umani e robot</b>   | <b>226</b> |
| 4.1     | <b>Intenzionalità: il marchio del mentale</b>                                                               | <b>228</b> |
| 4.1.1   | Tre approcci all'intenzionalità nelle scienze cognitive                                                     | 231        |

|                   |                                                                                                               |            |
|-------------------|---------------------------------------------------------------------------------------------------------------|------------|
| 4.1.1.1           | <i>Neo-Cartesianesimo: la visione internalista classica</i>                                                   | 234        |
| 4.1.1.2           | <i>Neo-Comportamentismo: strategie di previsione</i>                                                          | 237        |
| 4.1.1.3           | <i>Neo-Pragmatismo: esternalismo e norme sociali</i>                                                          | 245        |
| 4.1.1.4           | <i>Differenze e compatibilità</i>                                                                             | 248        |
| 4.1.2             | Intenzionalità e cognizione sociale: l'approccio classico basato sull'attribuzione                            | 251        |
| <b>4.2</b>        | <b>La mente che non c'è: attribuzione di intenzionalità negli HRI</b>                                         | <b>257</b> |
| 4.2.1             | L'importanza di essere "agente intenzionale"                                                                  | 257        |
| 4.2.2             | La misura dell'intenzionalità                                                                                 | 263        |
| 4.2.3             | Vecchie e nuove ibridazioni                                                                                   | 276        |
| <b>4.3</b>        | <b>Dall'attribuzione all'individuazione: un approccio enattivista per comprendere la HRI</b>                  | <b>283</b> |
| 4.3.1             | L'intenzionalità enattivista                                                                                  | 283        |
| 4.3.2             | Intentionality Detection: forme di riconoscimento dell'intenzionalità                                         | 293        |
| 4.3.3             | Un'analisi dell'intentionality detection nella HRI                                                            | 304        |
| <b>4.4</b>        | <b>La chiave dell'intenzionalità: il primato dell'<i>engagement</i></b>                                       | <b>320</b> |
| 4.4.1             | «Intentions in every case come already clothed in agency»: l'intenzionalità enattiva agli occhi di chi guarda | 323        |
| 4.4.2             | Il ritmo del vivente: evidenze dall'implementazione del movimento biologico                                   | 330        |
| 4.4.3             | Forme di intenzionalità sincretica: fuori dai correlati ontologici                                            | 335        |
| <b>Parte III</b>  | <b>Balla coi robot. L'alterità della macchina</b>                                                             | <b>343</b> |
| <b>Capitolo V</b> | <b>Il gioco dell'interazione. Rotture, abiti e aspettative nella HRI</b>                                      | <b>344</b> |
| <b>5.1</b>        | <b>Tra imitazione e interazione: l'enigma dei robot sociali</b>                                               | <b>344</b> |
| 5.1.1             | Il valore della rottura: per un'analisi dei breakdown                                                         | 348        |
| <b>5.2</b>        | <b>L'Uncanny Valley Effect</b>                                                                                | <b>354</b> |
| 5.2.1             | L'effetto perturbante: identificazione di uno strano fenomeno e delle sue possibili cause                     | 354        |
| 5.2.1.1           | <i>Spiegazioni percettive</i>                                                                                 | 364        |
| 5.2.1.2           | <i>Spiegazioni cognitive</i>                                                                                  | 368        |
| 5.2.2             | Sull'impostazione teorica e metodologica degli studi sull'UVE                                                 | 379        |
| 5.2.2.1           | <i>Note critiche al concetto di human-likeness</i>                                                            | 387        |
| 5.2.3             | Percezione è azione, azione è cognizione: significati incarnati e pratiche percettive                         | 392        |
| <b>5.3</b>        | <b>Interazioni di cristallo: coordinazione, abiti e aspettative nella HRI</b>                                 | <b>403</b> |
| 5.3.1             | Predictive Engagement: abiti e aspettative in interazione                                                     | 403        |

|            |                                                                                                                      |            |
|------------|----------------------------------------------------------------------------------------------------------------------|------------|
| 5.3.2      | Participatory Sense-Making, o la struttura dell'interazione sociale                                                  | 414        |
| 5.3.3      | L'UVE alla prova del gioco dell'interazione: la rottura dell'aspetto sincretico                                      | 427        |
| 5.3.3.1    | <i>Giochi a confronto</i>                                                                                            | 430        |
| 5.3.3.2    | <i>La danza dell'interazione: analisi dell'UVE agli occhi di chi guarda</i>                                          | 432        |
| 5.3.3.3    | <i>Esiti in itinere: uncanniness, micromondi enattivi e il robot come monstrum</i>                                   | 444        |
| <b>5.4</b> | <b>Galatea Futura</b>                                                                                                | <b>452</b> |
| 5.4.1      | Metodo Sintetico Enattivista                                                                                         | 453        |
| 5.4.2      | Kintsugi per l'HRI, o l'arte di tenere insieme i pezzi (dell'interazione): il caso della Learning from Demonstration | 457        |
|            | <b>Conclusioni e rilanci</b>                                                                                         | <b>467</b> |
|            | <b>Bibliografia</b>                                                                                                  | <b>472</b> |
|            | <b>Filmografia</b>                                                                                                   | <b>543</b> |



## Introduzione

Machines are becoming more human, so to speak – at least in the sense that, as Wiener indicated, some meaningful comparison exists between human and mechanical behavior.

But is it not ourselves that we know first and foremost? Rather than learning about ourselves by studying our constructs, *perhaps we should make the attempt to comprehend what our constructs are up to by looking into what we ourselves are up to.*

Perhaps, really, what we are seeing is a gradual merging of the general nature of human activity and function into the activity and function of what we humans have built and surrounded ourselves with.

Philip K. Dick (1973),  
*The Android and the Human*

Nel 1972, lo scrittore di fantascienza statunitense Philip Dick fu invitato a tenere un discorso alla University of British Columbia a Vancouver. Il discorso, intitolato *The Android and The Human*, cercava di sintetizzare le idee che lo scrittore aveva sviluppato nei suoi romanzi e racconti nel corso della sua intera carriera letteraria<sup>1</sup>, soprattutto riguardo ciò che era stato il tema al centro di tutti i suoi mondi fantascientifici fino ad allora, e che lo sarebbe stato anche per i suoi lavori futuri, ovvero la tensione primordiale che si instaura tra l'umano e la sua creatura artificiale. In particolare, quello che interessava allo scrittore era articolare la paura che l'umano, nel rapporto stretto che nei mondi distopici di Dick si forma con gli androidi, possa essere scambiato, prima o poi, per uno di loro.

Nei romanzi di Dick, l'androide è indistinguibile dall'essere umano, *sembra* a tutti gli effetti una persona sia a livello estetico, sia nei gesti, sia nei discorsi. È a partire, quindi, da questa somiglianza approssimata all'indistinzione che la paura di essere sostituiti si sposta su un altro livello, più profondo, e prende la struttura del paragone tra forme di empatia e comportamento, come accade nel suo romanzo più famoso *Do androids dream of electric sheeps?* (Dick 1968). Dal momento che, nei mondi di Dick, non c'è più possibilità di riconoscere gli androidi dalle loro sembianze, come è possibile distinguerli dagli altri che fanno parte della nostra stessa specie? Come fuggire dall'inganno che viene ordito da queste nuove forme di vitalità, che gioca proprio su cosa ci contraddistingue come esseri umani? Cosa ci rende umani, rispetto ai robot?

---

<sup>1</sup> In una lettera che accompagnava la versione scritta di questo discorso, pubblicato nel 1973 dalla rivista *SF Commentary* (Dick 1973), Dick concludeva: "this speech [tried]... to sum up an entire lifetime of developing thought".



Nel 1972, quando Dick teneva questo discorso, robot che avrebbero potuto presentare questo tipo di sfide al genere umano non erano realizzabili nemmeno all'interno dei più sviluppati laboratori di informatica. L'esempio più avanzato di tecnologia robotica disponibile in quel momento era SHAKEY, un robot appena costruito alla Stanford University che aveva l'obiettivo di muoversi nel proprio ambiente evitando gli ostacoli e spostando alcuni oggetti che incontrava sul proprio percorso. Le sue fattezze erano quelle di un carrello, con delle ruote che gli permettevano di muoversi, un grande apparato meccanico che dava forma al suo corpo e una lunga antenna tramite cui riceveva il segnale che proveniva da un computer situato nella stanza attigua. Nonostante SHAKEY incarnasse il traguardo più ambizioso raggiunto dalla tecnologia informatica, riusciva a fare ben poco: per percorrere qualche metro il robot impiegava ore, muovendosi lentamente a strattoni e fermandosi ogni qual volta riscontrava il minimo errore nel proprio funzionamento; nella maggior parte dei casi, falliva nel portare a termine i compiti che gli erano stati impartiti; la sua andatura traballante e il rumore costante emesso dalle sue componenti mentre era in funzione gli sono valsi il celebre soprannome. Oltre a SHAKEY, pochi altri esemplari di macchine autonome erano stati realizzati fino a quel momento; nessuna di queste poteva, nemmeno per errore, essere scambiata per un essere umano.

Perché, allora, Dick sentiva così pressante il timore di essere scambiato per un robot? Perché poneva al centro il tema della paura della *sostituzione* degli esseri umani da parte dei robot, anche se nella realtà questo pericolo era ancora così lontano? I temi del doppio, della creatura che si ribella al proprio creatore, della macchina che prende coscienza e che si vendica di chi l'ha costruita, sono stati motori della letteratura fantastica fin dalla tradizione mitologica, ma in quel periodo storico, negli ultimi quaranta anni del XX secolo, avevano ritrovato una nuova spinta grazie alle tecnologie realizzate a partire dagli anni '30, soprattutto quelle legate all'invenzione del computer e dell'Intelligenza Artificiale (IA). La fantascienza legata a questo tipo di macchine, che iniziavano ad apparire promettenti nel processo di comprendere il pensiero e le sue logiche tramite la riproduzione dei meccanismi interni, ha rilanciato la fantasia ancestrale connessa alla possibilità di scoprire, finalmente, l'essenza del nostro essere umani, e lo ha fatto soprattutto incarnandosi nella figura del robot: insieme a Dick, scrittori come Isaac Asimov e Arthur C. Clarke hanno contribuito in modo sostanziale alla diffusione dell'idea che macchine dalle intelligenze e dalle capacità sovraumane potessero confondersi nella nostra società e abitare il nostro stesso mondo, finendo per svelarsi come minacce per l'intero genere umano.

Non è un caso, allora, che oggi questa idea sia ancora più presente all'interno del nostro immaginario collettivo, modellando profondamente il modo in cui pensiamo ai robot sia nelle nostre vite quotidiane, sia negli ambienti accademici e scientifici. Al momento in cui viene consegnata

questa tesi, i traguardi dell'IA e della robotica sono arrivati a toccare frontiere che, negli anni '70, avrebbero fatto tremare le vene e i polsi: i nuovi modelli di intelligenza artificiale generativa – come i Large Language Model (LLM) *GPT-4*, *Claude*, *Gemini*, *LLaMA 2* – hanno rivoluzionato il modo di pensare all'IA e sembrano aver instradato la ricerca sulle reti neurali e il *deep learning* per gli anni a venire; generatori di immagini artificiali – come *DALL·E 3*, *Midjourney* e *Stable Diffusion* – stanno condizionando profondamente il nostro modo di considerare il rapporto tra le immagini e la realtà; di recente, la parata dei robot *Optimus Gen 2* prodotti da Tesla e i nuovi robot autonomi *Figure 01* di OpenAI hanno mostrato capacità straordinarie, mai raggiunte prima da alcuna macchina: danzano, lavano i piatti, eseguono istruzioni e – contemporaneamente – conversano con un interlocutore umano.

È costituendo un riflesso della nostra società ibrida, popolata in misura sempre maggiore da forme di intelligenza non organica, che anche i nostri racconti fantascientifici continuano a rinnovare il tema della *sostituzione* dell'essere umano da parte di macchine costruite allo scopo di incarnarne le forme umane del pensiero. In una diretta discendenza dalla letteratura al cinema e alle serie tv, racconti come quelli di Dick – e a volte proprio quelli, come testimonia il sequel del film *Blade Runner* (1982), *Blade Runner 2049* (2017) – ci prospettano una realtà alternativa o non troppo lontana in cui saremo rimpiazzati, sfidati, o a volte affiancati, da controparti robotiche: dal robot ginoide *Ava* in *Ex Machina* agli androidi ribelli di *Westworld*, dal *Terminator* dalla forza sovraumana al robot Lenny in *Io, robot*, questi racconti non fanno altro che costruire un immaginario che, al compimento dell'impresa di conoscenza che attualmente si sta svolgendo attraverso questo tipo di macchine intelligenti, ci ritroveremo improvvisamente ad abitare. In sostanza, quando si scoprirà la chiave che lega la macchina all'umano, quando tutti i tasselli che compongono i meccanismi dell'intelligenza saranno messi al proprio posto, sarà il momento in cui arriveremo finalmente a comprendere ciò che ci rende umani e, allo stesso tempo, quello in cui avremo perso l'esclusiva. Il risultato è che, in questa prospettiva, quello che ci contraddistingue ancora come esseri umani in quanto creature intelligenti consiste in tutto ciò che non è ancora stato formalizzato in una macchina<sup>2</sup>.

Questo principio, lungi dall'essere solamente un pensiero formulato all'interno di racconti fantascientifici e mondi distopici, caratterizza la ricerca nel campo dell'IA e della robotica fin dalla loro fondazione. In particolare, l'IA nasce con il proposito programmatico secondo cui «every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it» (McCarthy et al. 1955 (2006): 12). Gli immediati salti in avanti compiuti dalla nuova disciplina nei primi anni del suo esercizio, con risultati come il *Logic Theorist*

---

<sup>2</sup> Come definito da Douglas Hofstadter nel teorema di Tesler: «L'IA è tutto ciò che non è ancora stato fatto». (Hofstadter 1979: 649)

(LT, Newell & Simon 1956), il *General Problem Solver* (GPS, Newell et al. 1959) o *Eliza* (Weizenbaum 1966), avevano effettivamente fornito delle basi che apparivano solide per esaudire la speranza di riuscire finalmente a comprendere il modo in cui funziona il nostro pensiero. Sulla scia di queste conquiste tecnologiche, anche le neonate scienze cognitive, costituite allo scopo di studiare la mente umana attraverso un approccio interdisciplinare – che comprendeva, tra le altre, psicologia, teoria dell'informazione, linguistica e filosofia (Dupuy 1994; Boden 2006) – trovarono nell'IA e nelle macchine che la implementavano un alleato metodologico privilegiato. L'idea che i processi cognitivi potessero essere formalizzati in termini computazionali e implementati in macchine aprì così nuove prospettive di ricerca, formalizzando una metafora che segna uno dei principi fondativi cardine delle scienze cognitive classiche e, per certi versi, anche di alcune influenti varianti contemporanee, ovvero quella della mente come elaboratore di informazioni (Block 1993, 1995; Gigerenzer & Goldstein 1996).

In questo quadro, il ruolo delle macchine intelligenti, robot compresi, consiste nel riprodurre fedelmente quello che si crede essere il funzionamento del pensiero: è in questo punto che si aggancia il tema fantascientifico della replicazione dell'umano, poiché macchine che possono funzionare esattamente nel modo in cui noi esseri umani funzioniamo non impiegheranno molto tempo a prendere il nostro posto, a insinuarsi nelle nostre attività sociali, a svolgere i lavori che ad oggi risultano ancora nostra prerogativa, magari a sorpassarci in intelligenza o capacità di ragionamento. Così, il mito della sostituzione si rinnova e trova terreno fertile, ancora oggi, per popolare le nostre speranze nel caso dei tecnoentusiasti, o le nostre paure nel caso dei più apocalittici. Di recente, il premio Nobel 2024 per la fisica Geoffrey Hinton, celebre per aver inventato negli anni '80 gli algoritmi di *backpropagation* alla base del *machine learning* (LeCun et al. 1988; Lillicrap et al. 2020), ha pubblicamente affermato che con il proprio lavoro ha dato via a un percorso irreversibile in cui le macchine non solo sapranno fare quello che noi sappiamo fare, ma anche molto meglio: «in trying to mimic what biological brains do, we've come up with something better» (Hao 2023).

Se, da una parte, assistere alle stupefacenti performance linguistiche e analitiche di GPT-4 potrebbe suscitare in noi le stesse impressioni di Hinton, dall'altra, però, basta osservare le prestazioni dei robot di oggi per recuperare un sano scetticismo. Infatti, quello che spesso viene messo in secondo piano è che, come accadeva ai tempi di Dick, anche oggi se guardiamo all'interno dei laboratori più avanzati non troviamo robot molto più capaci di SHAKEY di operare in ambienti non controllati, e tanto meno di sostituirci pienamente nell'esercizio dei nostri compiti quotidiani. I robot *Optimus Gen 2* e *Figure 01* di cui sopra hanno, sì, mostrato capacità sorprendenti, ma non erano completamente autonomi: mentre il primo era dichiaratamente teleoperato a distanza da un essere umano che gli inviava comandi permettendogli di muoversi in modo agile e dialogare fluidamente con i passanti, le

performance esibite dal secondo nel video di presentazione erano il risultato di un addestramento mirato a svolgere quelle azioni. Quello che si vuole sottolineare è che, nonostante gli avanzamenti in merito di tecnologia robotica e intelligenza artificiale siano stati giganteschi soprattutto negli ultimi dieci anni, l'obiettivo di raggiungere prestazioni anche solamente paragonabili in termini di autonomia, spontaneità e realizzazione a quelle degli esseri umani resta ancora uno dei sogni a lungo termini degli ingegneri informatici e robotici.

L'esempio più adatto a descrivere questo paradosso, in cui si articola la complessità delle tecnologie robotiche ad oggi a disposizione e l'impossibilità di raggiungerne di sufficientemente tali per costruire artefatti che possano eguagliare le nostre capacità di interazione con il mondo è il robot androide *Geminoid HI-2*, creato a immagine e somiglianza dall'ingegnere robotico Hiroshi Ishiguro. *Geminoid* potrebbe apparire potenzialmente capace di svolgere qualsiasi dei compiti che normalmente svolge il suo padre ingegnere, a cui somiglia così tanto: ha due braccia per afferrare oggetti, un volto che esprime emozioni, con occhi per guardare, una bocca per parlare, e gambe che gli consentirebbero di spostarsi ovunque esso voglia. Il punto, però, è che *Geminoid* non ha alcuna autonomia e viene utilizzato da Hishiguro come una protesi, che spedisce per posta nei posti in cui deve tenere conferenze e che comanda a distanza, dando l'impressione che sia effettivamente il robot a compiere quei gesti e a tenere quei discorsi. L'effetto di realtà è notevole, ma *Geminoid* non è altro che un sofisticato strumento di teleoperazione: le tecnologie robotiche gli hanno consentito di generare movimenti minimi quando in funzione, di compiere gesti ripetitivi come battere le ciglia o indirizzare il proprio volto verso una sorgente di rumore, ma ha molto di più in comune con un burattino che con una macchina automatica.

Nonostante i decenni di ricerca che separano i robot di oggi dai primi esemplari degli anni '70, infatti, la maggior parte dei robot contemporanei rimangono limitati e confinati nell'esecuzione di esercizi elementari: afferrare un oggetto, spostarlo, o camminare. Più che emulare il nerboruto *Terminator* o la sofisticata Dolores di *Westworld*, i robot di oggi sono più vicini a *Emilio*, il robottino giocattolo che ha entusiasmato l'infanzia di molti bambini nati negli anni '90, promettendo agognati quanto illusori servizi domestici, inseguimenti sfrenati, giochi condivisi. Molte delle creature che costruiamo hanno ancora enormi problemi nell'esecuzione di compiti che all'apparenza potrebbero sembrare elementari, quasi banali, ma in cui in realtà risiede lo scoglio più grande all'effettiva costruzione di artefatti capaci di esibire livelli accettabili di intelligenza pratica: renderli capaci di muoversi con naturalezza nel nostro stesso mondo.

È cruciale notare che un notevole passo in avanti verso il raggiungimento di questo obiettivo è stato compiuto a partire dagli anni '80, quando in robotica hanno acquisito un particolare rilievo fattori che sino a quel momento erano stati lasciati da parte nella descrizione della cognizione e della

sua modellizzazione. Considerare fattori come il corpo, l'apparato sensomotorio, il rapporto dell'agente con il proprio ambiente e le dinamiche situate del comportamento ha avuto come risultati immediati quelli di realizzare robot sempre più adattivi e capaci addirittura di intrattenere delle interazioni di tipo sociale con gli esseri umani (Breazeal et al. 2016). È da tale possibilità che si è sviluppato un intero campo di indagine dedicato alla *Human-Robot Interaction* (HRI) (Dautenhahn 2007, 2014). Questo campo di ricerca nasce esattamente al crocevia tra scienze dell'artificiale e scienze cognitive, terreno comune in cui si incontrano intelligenze umane e cibernetiche, portando all'attenzione di entrambi i domini la necessità di analizzare l'umano e il robot non solo separatamente, ma nella loro relazione, anche e soprattutto se l'intento è quello di costruire artefatti che sappiano come interagire con noi. Infatti, negli anni questa indagine ha contribuito in modo sostanziale a nutrire le ricerche nel campo della robotica sociale, instradando la costruzione di robot le cui caratteristiche e capacità venivano identificate da uno studio focalizzato alla comprensione della socialità e delle dinamiche relazionali nell'essere umano (cfr. §3.2).

Parallelamente, a partire dagli anni '90 e in concomitanza con i risultati ottenuti dalla robotica situata, una nuova proposta si è fatta strada come alternativa alle prospettive cognitiviste elaborate nei primi quaranta anni delle scienze della mente, basata sui principi fondativi dell'*embodiment* e della cognizione situata: l'enattivismo (Varela et al. 1991; cfr. §2.3.3). Il principale obiettivo di questa corrente di pensiero all'interno delle scienze cognitive può essere rintracciato nell'intento di disimplicare l'indagine della mente dalla metafora computazionale e fondare la comprensione dei processi cognitivi sulla considerazione e sull'analisi delle interdipendenze tra corpo, cervello e ambiente che soggiacciono all'emersione di forme organiche di comprensione del mondo. Nei trenta anni che sono intercorsi dalla ufficiale definizione del paradigma enattivista ed *embodied*, le scienze cognitive hanno subito una ristrutturazione anche a causa delle ingenti influenze apportate da questa nuova apertura teorica (cfr. Chemero 2009; Gallagher 2023a; Stewart et al. 2010). Molti temi di ricerca hanno subito una rilettura grazie all'intervento delle teorie enattiviste di mente e cognizione, sviluppando riflessioni in merito alle basi biologiche e formali della cognizione (Maturana & Varela 1980; Varela et al. 1991), alla percezione (Noë 2004; Degenaar & O'Regan 2017), all'agentività (Di Paolo et al. 2017), al linguaggio (Di Paolo et al. 2018), all'intersoggettività e alla cognizione sociale (Gallagher 2021; De Jaegher & Di Paolo 2007), al sé (Zahavi 2014), alle psicopatologie (Lobaccaro 2022a; Colombetti 2013) e altro ancora.

Tuttavia, ad oggi l'indagine sull'interazione tra umani e robot non è ancora stata oggetto di alcuno studio da questo punto di vista. Il paradigma attraverso cui la maggior parte delle indagini sulla HRI fanno riferimento è quello legato alla prospettiva cognitivista e, di conseguenza, si lega saldamente al tema della sostituzione, esattamente come lo abbiamo descritto fino a qui: i robot devono essere

costruiti implementando le ipotesi che le scienze della mente forniscono riguardo al funzionamento della cognizione umana individuale, con la speranza di trovare l'esatta formula che possa posizionare correttamente ogni tassello e replicare in questo modo le forme di intelligenza che sottoponiamo a indagine. Come abbiamo visto, però, il perseguimento del progetto di ricerca basato sulla metafora computazionale non ha consentito, almeno fino ad ora, di realizzare forme complesse di artefatti robotici che sappiano interagire efficacemente e fluidamente nel nostro mondo. Anche i robot sociali più all'avanguardia, come *Sophia* (Riccio 2024), *iCub* (Metta et al. 2010) o *Nao* (Shamsuddin et al. 2011), non fanno eccezione: riescono a compiere azioni straordinarie – come riconoscere la direzionalità dello sguardo e l'intonazione della voce, leggere le emozioni, rispondere a qualche domanda – ma, alla stregua dei robot più avanzati di cui abbiamo parlato, non riescono a imbastire interazioni complesse, né a sostenere la dinamicità dell'incontro con un essere umano in situazioni di vita quotidiana. Tramite l'implementazione delle più recenti tecnologie di IA nelle architetture robotiche, come il *machine learning*, i Big Data o l'IA generativa, si è riusciti a costruire macchine ancora più complesse, sicuramente migliorate nelle performance interazionali, ma ancora lontane dall'essere capaci di sostenere come noi e con noi interazioni sociali complesse. Anche oggi, dunque, il momento della sostituzione sembra essere qualcosa di riservato agli abitanti di un futuro lontano, quello in cui si sarà avverata la profezia della metafora computazionale.

L'obiettivo principale di questa tesi consiste nell'introdurre, nel novero delle teorie che nutrono gli studi di HRI e che informano la progettazione di robot sociali, una prospettiva alternativa come quella enattivista: un tentativo che è ancora inedito nel panorama attuale e che, questa la nostra ipotesi di partenza, crediamo renderà possibile fornire una rilettura del modo in cui concepiamo i robot, sia nel processo di progettazione e costruzione, sia nelle modalità di analisi e comprensione dell'interazione che intraprendiamo con loro. La proposta si fonda sull'osservazione delle precedenti spinte operate dalle scienze cognitive enattiviste nel campo della robotica, che già hanno contribuito almeno una volta – nel caso della *Nouvelle AI* (§2.3.2) – a infondere idee che si sono rivelate preziose per lo sviluppo di artefatti complessi e capaci di agire in modo più adattivo nel proprio ambiente.

Parleremo estensivamente in questa tesi delle possibilità aperte da questa visione in alternativa alle teorie cognitiviste, proponendo di volta in volta un ribaltamento di prospettive rispetto a temi specifici. Ne tratteremo soprattutto per specificare il modo attraverso cui sia possibile costruire un approccio allo studio dell'interazione tra umani e robot che tenga in considerazione il ruolo del corpo, l'incontro incarnato, le qualità sensomotorie dell'esperienza di fronte al robot, le caratteristiche della intrinseca socialità della cognizione umana, la peculiarità dell'interazione sociale che si instaura e le funzioni che l'ambiente, il contesto e le dinamiche situate svolgono all'interno di questa. In questo momento, ciò che ci preme sottolineare è la possibilità aperta da questo cambiamento epistemologico,

che ci consente lo sviluppo di nuovi modelli per comprendere il rapporto che si instaura in interazione con le tecnologie robotiche; modelli che, auspichiamo, possano prospettarci nuovi orizzonti di collaborazione con i robot sia sul piano pratico, sia teorico.

### *Verso nuovi modelli per l'interazione tra umani e robot*

Il primo passaggio di modelli che vogliamo proporre in questa tesi è quello che riguarda la rielaborazione del paradigma della sostituzione, al fine di creare un modello di *collaborazione* fra umani e robot. Strettamente legato al modello cognitivista che ha costituito la cifra delle scienze cognitive classiche, il modello della sostituzione ancora oggi è quello che più frequentemente si ritrova all'interno degli studi di HRI. La prospettiva enattivista ci consentirà di sganciare lo studio della mente dalla metafora computazionale e di operare un aggiornamento del modo in cui è possibile guardare alle macchine per comprendere la struttura della nostra mente: pensare che macchine come i robot possano ricoprire esclusivamente il ruolo di riproduzioni dell'essere umano preclude la possibilità di instaurare una relazione di collaborazione in cui entrambi gli agenti riescano a contribuire effettivamente all'interazione.

Operare uno spostamento di prospettive rivolto a costruire un modello di collaborazione rende possibile un ripensamento del modo in cui le unità che compongono la HRI vengono indagate. Nel dominio disciplinare della HRI, i modelli di cognizione sociale adottati nella maggior parte degli studi implicano una prospettiva individuale dell'unità implicata nell'interazione. In altre parole, l'interazione è considerata come la somma delle azioni e delle reazioni che ciascuna delle parti rivolge verso l'altra. Ciò comporta la perdita della specificità che contraddistingue l'interazione tra umani e robot, resa invece esplicita se si adotta uno sguardo enattivista ed *embodied*. Adottando questo approccio, infatti, il percorso che viene reso possibile consiste nel passaggio da un modello che considera l'apporto dell'individuo in interazione a uno che considera l'interazione come il *dialogo* situato e incarnato tra le parti.

Il prospetto delineato dai modelli della collaborazione e del dialogo ci consentirà un'ultima operazione teorica, che consiste nel passaggio da un modello della replicazione a quello dell'*alterità*. In questo modello, ciò che si vuole portare all'attenzione è la possibilità di considerare il robot come un partner attivo in interazione che, interrogato, possa restituirci risposte inaspettate alle domande poste sulla natura della nostra capacità interazionale. Adottando questo modello, potremo esplicitare i vantaggi di considerare la HRI come il luogo in cui si incontrano diverse forme di cognizione, umana e artificiale, radicalmente differenti ma non necessariamente impossibilitate a comunicare.

### *L'oggetto di indagine*

L'oggetto al centro di questa indagine è costituito di tre parti diverse, ma interconnesse: il robot, l'umano e la loro interazione. Nel corso della trattazione, cercheremo di mostrare come la ricerca sull'interazione non possa prescindere da una disamina approfondita delle unità dialoganti che ne permettono la realizzazione: per questo, il robot sarà analizzato nelle sue componenti mimetiche e sintetiche, che ne permettono l'emersione come figura sincretica in interazione agli occhi dell'interlocutore; l'umano, d'altra parte, sarà preso in considerazione attraverso le descrizioni fornite dagli approcci enattivisti alla cognizione sociale, attraverso cui verrà resa intelligibile la struttura partecipata che permette l'allestimento di configurazioni di senso in interazione; quest'ultima sarà considerata come il frutto dinamico dell'incontro in atto, il quale a sua volta presenta la capacità di indirizzare le azioni e la partecipazione degli attori coinvolti nel flusso condiviso della reciprocità.

### *Il campo di studi*

Nella composizione interdisciplinare di questa tesi, il dominio di studi di *Human-Robot Interaction* costituisce il campo specifico di posizionamento dell'indagine. La vocazione teorica e filosofica di questo lavoro non resta immune alle possibili applicazioni pratiche che possono informare una riconfigurazione degli studi di settore a partire dalla considerazione di idee e ipotesi differenti sul funzionamento della mente e sulle modalità di costruire artefatti robotici che abbiano lo scopo di interagire efficacemente con l'essere umano. Inserire in questo dominio le ipotesi sulla dinamicità e sulla struttura incarnata della mente consentirà di aprire uno spazio di indagine più ampio di quello attualmente delineato negli studi di HRI, in cui il robot verrà considerato attraverso i modelli della collaborazione, del dialogo e dell'alterità. L'approccio che adotteremo si posiziona all'interno di una prospettiva *human-centered*, ovvero che tiene in considerazione prima di tutto le esigenze e le caratteristiche del coinvolgimento umano nell'interazione con il robot.

La *filosofia*, in questo quadro, interviene a formare legami e ad aprire confronti, svolgendo il ruolo di ponte tra le discipline e traducendo laddove possibile le compatibilità degli approcci all'interno dello stesso campo di ricerca e tra indagini differenti. In un contesto in cui scienze cognitive, robotica e studi di HRI tendono spesso a rimanere ancorati a paradigmi operativi distinti, la filosofia agisce come uno spazio di sintesi e di interrogazione sulle basi epistemologiche e ontologiche di tali approcci, esplorando le potenzialità di un'autentica interazione tra umano e robot. Attraverso il dialogo tra discipline realizzato sotto la guida della riflessione filosofica, la tesi esplora la possibilità di superare l'individualismo metodologico che spesso caratterizza le scienze cognitive, valorizzando invece un concetto di cognizione come pratica condivisa e come processo intrinsecamente legato alla relazione tra corpo e ambiente. Il contributo filosofico consiste, quindi, nell'armonizzare e rendere



compatibili le varie prospettive, valorizzando le potenzialità di ogni metodo per sviluppare una visione integrata e dialogica dell'interazione tra umani e robot.

È adottando una prospettiva così composita e interdisciplinare, guidata dalle ipotesi enattiviste sulla struttura della mente e della cognizione umana, che sarà possibile affrontare alcune delle maggiori sfide individuate nello stato dell'arte della HRI. In particolare, l'antropomorfismo, l'*embodiment*, l'attribuzione di intenzionalità e l'*Uncanny Valley Effect* costituiscono i fenomeni al centro del dibattito contemporaneo negli studi sull'interazione tra umani e robot, che al momento rimangono insondati attraverso uno sguardo che consideri il ruolo del corpo e dell'interazione dinamica tra agente e ambiente. Il risultato che si prospetta è quello di raggiungere un quadro unitario che sia capace di dare conto dei fenomeni in modo coerente, elaborando metodi di analisi che cerchino di tenere in considerazione le molteplici variabili di cui si compone la struttura della cognizione umana.

### *Le tappe del percorso: la struttura dell'elaborato*

L'elaborato si articola in cinque capitoli, raccolti in tre parti principali che costituiscono i passaggi principali del nostro ragionamento.

Nella prima parte, *Storia ed epistemologia della robotica. Sullo sviluppo delle tecniche di mimesi e riproduzione dell'umano*, vengono presi in esame i concetti e le condizioni storico-filosofiche che hanno reso possibile la nascita del robot come entità capace di implementare teorie della cognizione umana. La sezione esplora le radici epistemologiche e tecnologiche della robotica, concentrandosi sugli sviluppi teorici nel campo delle scienze cognitive, a partire dagli anni '60. In particolare, la ricostruzione critica si articola in due capitoli:

- *Capitolo I – Macchine come me. L'ambizione dell'umano come conoscenza attraverso la macchina.* In questo capitolo, si ricostruisce la genesi del robot come macchina cibernetico computazionale a partire da un'analisi dell'ambizione dell'umano e degli sviluppi del concetto di macchina dall'antichità fino all'avvento dell'intelligenza artificiale. Attraversando diverse fasi di evoluzione del concetto di macchina, verrà fatto notare come l'ambizione dell'umano di costruire artefatti che abbiano capacità agitive abbia condotto nel corso dei secoli ad una stratificazione del concetto di macchina, arrivando poi a gettare le base per la costituzione di una disciplina dedita alla realizzazione di macchine capaci di esibire un comportamento: la robotica. Passaggi obbligati in questo percorso sono rappresentati dalle teorie meccaniciste cartesiane, dalla cibernetica come disciplina che raccoglie la sfida di coniugare indagini sulla vita e realizzazione di artefatti artificiali e dalla realizzazione dell'intelligenza artificiale come presupposto alla costituzione della robotica.

- *Capitolo II – Dalle scienze cognitive alla robotica: andata e ritorno.* Il percorso realizzato nel capitolo precedente consente un'analisi approfondita delle condizioni di realizzazione dei primi robot autonomi, che tenga in considerazione le ipotesi sul funzionamento della mente elaborate dalle allora neonate scienze cognitive e della loro traduzione nel campo delle scienze dell'artificiale. Nel viaggio di andata, dunque, si fornisce un'analisi del primo robot autonomo, SHAKEY, mettendo in luce le ipotesi cognitive che ne hanno informato la costruzione. In seguito, a partire da una serie di critiche formulate sulla qualità delle prestazioni del robot, si procede in un viaggio di ritorno dalla robotica alle scienze cognitive, innescato da una serie di osservazioni relative a tesi alternative sulla costituzione della mente e della cognizione sviluppate nell'ambito del movimento della seconda cibernetica, dell'intelligenza artificiale e dell'approccio filosofico a questi temi. Il risultato porterà a riconsiderare il ruolo delle teorie su mente e cognizione nel processo di costruzione dei robot con vocazione sociale, rendendo possibile una lettura alternativa delle possibilità incarnate dalle macchine nella comprensione della cognizione e dell'esperienza umana.

La seconda parte, *Ti estì. Il robot agli occhi dell'osservatore*, è dedicata all'analisi del robot sociale all'interno degli studi sull'interazione tra umani e robot. In particolare, viene fornito lo stato dell'arte in merito a fenomeni centrali nella HRI – quali cognizione sociale, antropomorfismo, *embodiment* e attribuzione di intenzionalità – e viene dimostrata la necessità di un cambiamento delle teorie sulla cognizione e sulla mente che informano l'analisi dell'interazione tra umani e robot. Il percorso consiste nell'elaborazione della figura del robot come artefatto composto di tre aspetti: componente mimetica, componente sintetica, componente sincretica. L'analisi consta di due capitoli.

- *Capitolo III – Galatea moderna. Robot sociali tra modellizzazione e mimesi del vivente.* A partire da una disamina sullo stato dell'arte degli studi sulla HRI, si procede alla definizione dei problemi che identificano la disciplina e viene rintracciato un comune approccio riguardo alle questioni concernenti la cognizione sociale. In particolare, viene sottolineato come lo studio dei processi cognitivi volti alla comprensione dell'agente in interazione sia effettuato quasi esclusivamente attraverso gli strumenti concettuali forniti dalla teoria della mente (ToM). Vengono rintracciate le conseguenze di questo connubio sulla considerazione del robot in interazione e viene fornita un'alternativa che rende possibile una rivalutazione del ruolo del corpo, della socialità e del rapporto con l'ambiente nei processi cognitivi umani. Si procede, dunque, a sviluppare un apparato critico individuando due aspetti fondamentali mostrati dal robot in interazione: a) il robot si presenta come artefatto *mimetico*, ovvero capace di esibire livelli di somiglianza all'umano unici tra gli artefatti artificiali: tale aspetto viene analizzato nello studio del fenomeno di antropomorfismo nei confronti del robot; b) il robot si

contraddistingue tramite il suo aspetto *sintetico*, ovvero per la capacità di esibire un comportamento realizzato in un ambiente: tale aspetto viene rintracciato a partire da una disamina dello strumento concettuale dell'*embodiment* e dalla definizione della questione del comportamento. Viene mostrato, infine, come la cifra distintiva del robot sia quella di esibire questi due aspetti contemporaneamente nella presentazione agli occhi dell'interlocutore umano, ovvero la sua componente *sincretica*.

- *Capitolo IV – Umani, troppo umani. Intenzionalità e agentività per i robot sociali.* In questo capitolo viene fornita un'analisi approfondita del robot in quanto artefatto sincretico tramite la considerazione del fenomeno dell'attribuzione di intenzionalità da parte dell'essere umano. Rintracciando le radici filosofiche del concetto di intenzionalità, si procede all'analisi degli approcci maggiori nelle scienze cognitive e dei modi in cui l'attribuzione di intenzionalità viene indagata negli studi di HRI. Vengono dunque rilevate alcune inconsistenze epistemologiche e difficoltà interpretative nell'apparato sperimentale, per cui si propone una soluzione tramite l'adozione di un concetto enattivista di intenzionalità e della sua attribuzione ai robot sociali. In questo quadro, si profila la necessità di un cambiamento non solamente concettuale, ma anche nominale, e si propone la dicitura *intentionality detection* per sottolineare la rottura con gli approcci classici alla questione. Il robot viene quindi descritto esibire un'*intenzionalità sincretica*, non in quanto caratteristica intrinseca della sua forma di comprensione, ma come frutto dell'interazione sociale istituita con l'essere umano. Infine, questa lettura renderà possibile operare una ulteriore specificazione della nozione di intenzionalità enattivista promossa dall'applicazione nel campo della HRI, effettuata recuperando il rapporto originale che costituisce il nucleo dell'intenzionalità enattivista radicale: il legame con l'agentività del sistema cognitivo.

La terza e ultima parte, *Balla coi robot. L'alterità della macchina*, consiste nel tentativo di costruire un approccio enattivista allo studio dell'interazione tra umani e robot e a testare le ipotesi operative su uno dei fenomeni più complessi del settore degli HRI: l'*Uncanny Valley Effect* (UVE). Si propone, quindi, un'analisi del fenomeno a partire dalle considerazioni elaborate nei capitoli precedenti e fornendo un quadro dettagliato della struttura delle interazioni sociali attraverso l'impiego delle teorie del *predictive engagement* e del *participatory sense-making*. In conclusione, si delineano delle possibilità di indagine aperte dal cambiamento operato tramite la lettura enattivista dei fenomeni osservati, sia dal punto di vista teorico, sia delle applicazioni pratiche. La parte si costituisce di un capitolo.

- *Capitolo V – Il gioco dell'interazione. Rotture, abiti e aspettative in HRI.* Il capitolo procede a sviluppare una variante del gioco dell'imitazione proposto da Alan Turing, ovvero il gioco

dell'interazione. Il gioco dell'interazione si delinea attraverso la teoria dei micromondi elaborata da Francisco Varela, in cui assume un ruolo centrale il momento di *breakdown*. Per realizzare un'analisi della HRI attraverso il gioco dell'interazione, si prende quindi in esame la rottura interazionale dell'*Uncanny Valley Effect*. Dopo una disamina delle spiegazioni fornite nello stato dell'arte attuale per dare conto dell'effetto perturbante di fronte al robot, si procede con l'introduzione delle ipotesi enattiviste del *predictive engagement* e del *participatory sense-making* al fine di fornire strumenti teorici per spiegare il fenomeno *uncanny* tramite il rapporto che si costituisce nell'interazione sociale tra abiti e aspettative incarnate e concretezza dell'incontro con il robot. L'UVE viene spiegato nella struttura dei micromondi interazionali che costituiscono il gioco dell'interazione come l'affastellarsi di *micro-breakdown* che non vengono recuperati in modo congiunto tra umani e robot, che si situano principalmente a livello sensorimotorio e incarnato. In seguito, si prova ad aprire due linee di ricerca a partire dalle considerazioni elaborate nel capitolo in particolare, e nella tesi in generale. Sul piano teorico ed epistemologico, si propone la visione di un metodo sintetico enattivista in cui i sistemi *target* non sono più solamente individuati nella struttura interna e individuale del comportamento o della cognizione, ma nei processi di coordinazione e aggiustamento presupposti dalle interazioni sociali, rendendo così possibile basare la costruzione di robot sociali sui processi interazionali osservabili e non su strutture interne della mente. Il ruolo dell'interazione come elemento cardine di questo aspetto viene poi messo in relazione alle possibilità pragmatiche di intervento sulla progettazione delle architetture cognitive robotiche nel campo della *learning from demonstration*.

In conclusione a questa introduzione, e all'inizio di questo elaborato di tesi, vogliamo riassumere l'intento di questo lavoro cogliendo il senso della citazione di Philip K. Dick che abbiamo inserito all'inizio di questo paragrafo: un invito programmatico verso la possibilità di rileggere il nostro rapporto epistemico con i robot. Anche se nella storia delle scienze dell'artificiale e della mente si è guardato alle macchine principalmente per comprendere quanto queste siano simili a noi, Dick suggerisce di ribaltare la prospettiva e innanzitutto cercare di comprendere meglio noi stessi. Solo così saremo capaci di comprendere le macchine che costruiamo e così impiegarle per scavare ancora più a fondo nell'impresa di conoscere le forme della nostra capacità di comprendere il mondo e gli altri agenti che lo abitano. Le macchine che costruiamo non possono far altro che incorporare i principi che gli impartiamo, implementano inevitabilmente i nostri modi di pensare, le nostre logiche, i nostri obiettivi di ricerca e di applicazione. Con questa tesi, cercheremo di ampliare gli orizzonti attraverso cui i principi in questione vengono formulati.



**Parte I -**

**Storia ed epistemologia della robotica.**

**Sullo sviluppo delle tecniche di mimesi e riproduzione dell'umano**

## Capitolo I – Machine come me.

### L'ambizione dell'umano come conoscenza attraverso la macchina

Lifelike – but not alive. [...] So, what's the difference  
between my pain and yours? Between you and me?  
*Westworld* (2018-2022)

«Cosa mi rende diverso da te?» chiede il robot Bernard al suo costruttore, il dottor Ford, durante uno dei momenti più intensi della serie televisiva *Westworld* (2018-2022). In questa serie, un parco a tema popolato da robot senzienti funge da scenario per la realizzazione fantascientifica di un dilemma millenario, che affonda le proprie radici nella nascita della capacità dell'umano di immaginare e di costruire. Cosa ci rende diversi dalle nostre creature artificiali? La serie, capolavoro nato dalla fantasia di Jonathan Nolan e Lisa Joy, si snoda cercando di rispondere a questa domanda, inscenando dialoghi che portano distintamente il marchio di un ragionamento filosofico appartenente al pensiero dell'ultimo secolo sulla natura della mente, su cosa sia la coscienza<sup>3</sup>, la cognizione, e come sia possibile per l'umano realizzare un suo corrispettivo artificiale. Il tema del doppio, del clone, dell'imitatore, uno dei più grandi *leitmotiv* della cultura occidentale, prende qui le sembianze di Bernard: la capacità raggiunta dal robot di porre questioni esistenziali al suo costruttore ha il ruolo di mettere l'umano di fronte a ciò che, per definizione, non è, ma che gli somiglia terribilmente. Macchina al contempo amletica e ingannatrice, il robot Bernard innesca una riflessione che mette in crisi il dottor Ford, moderno dottor Frankenstein che deve necessariamente ricorrere a stratagemmi ed esperimenti mentali per individuare la differenza: in quel momento della serie, al robot Bernard manca la coscienza, non è libero di errare – vagare, ma anche sbagliare – nel labirinto dell'esistenza, e i suoi ricordi sono costruiti o impiantati artificialmente, rubati a persone in carne ed ossa che hanno forse davvero vissuto la propria vita, ora condensata e surrogata in qualche migliaio di bit. E tuttavia, tale escamotage non è sufficiente. In *Westworld*, il gioco delle parti sta proprio nel far vedere, attraverso la tecnica che ha permesso al genere umano di realizzare delle copie praticamente esatte di sé stesso, come ci sia una doppia faccia. Alla domanda posta da Bernard, il dottor Ford risponde:

The answer always seemed obvious to me. *There is no threshold that makes us greater than the sum of our parts.* No inflection point at which we become fully alive. We can't define consciousness because consciousness does not exist. Humans fancy that there's something special about the way we perceive the world and yet, we live in loops, as tight

---

<sup>3</sup> La serie prende esplicitamente spunto dalla riflessione sulla mente bicamerale elaborata dallo psicologo Julian Jaynes (1976).

and as closed as the Hosts do. Seldom questioning our choices, content for the most part to be told what to do next.

Non c'è soglia che ci renda qualcosa di più della somma delle nostre parti, che ci renda completamente vivi, dice Ford, la stessa coscienza non esiste. Sono concetti, come quello di libertà, altro filo conduttore di tutta la serie, che l'uomo ha costruito per identificarsi nella sua unicità, in assenza di dialogo con le creature che prendeva a contraltare della propria essenza. Ora che le nostre creature possono parlarci, ci mettono di fronte al dedalo delle nostre responsabilità, delle nostre azioni, del risultato di un percorso di sviluppo lungo quanto la storia della tecnica. La morale, in questo caso, è che il rapporto con le macchine antropomorfe può rivelarsi non solo nelle sue differenze, ma soprattutto nelle caratteristiche comuni.

Il dialogo tra il dottor Ford e Bernard mostra però un doppio dilemma, qui presentato come un'unica questione problematica, ma che si ritiene necessario disimpiacare. Da una parte, il dolore provato da Bernard, il fatto che sia in tutto e per tutto uguale al suo costruttore, che sia capace di sentimenti e di assumersi le responsabilità delle proprie azioni, mostra il raggiungimento di un obiettivo che l'umanità si è posta fin dall'antichità, ovvero quella di realizzare delle copie indistinte di sé stessa. Dall'altra parte, questa somiglianza approssimata all'indistinzione porta il dialogo su un altro livello: data la realizzazione del mito della copia, ora finalmente le macchine possono dirci qualcosa su cosa ci contraddistingue; dato che siamo uguali sul piano corporeo e condividiamo gli stessi meccanismi che generano le nostre illusioni, allora possiamo confrontarci per capire cosa rende noi umani e loro macchine.

È proprio da qui, allora, che vogliamo partire per il presente capitolo, affrontando criticamente questa doppia anima e sottolineando come la robotica, in quanto disciplina dedita alla realizzazione di artefatti artificiali utili e a volte somiglianti all'umano, sia l'ultimo passo di un percorso di immaginazione e realizzazione attraverso la tecnica. Da una parte, infatti, la robotica raccoglie l'ambizione dell'umano di costruire artefatti a lui simili, che possano fungere da sostituti funzionali per alleviare le fatiche del lavoro, per raggiungere traguardi tecnici, o per incarnare miti e leggende circa il potere di creare vita dalla materia e l'impossibilità di governarla fino in fondo. Dall'altra parte, nel processo di sviluppo di tecniche di imitazione, mimesi e riproduzione dell'umano, l'ambizione si trasforma e diventa ambizione di *comprendersi*. Infatti, a partire dalla metà del XVII secolo, le macchine non vengono prese solamente a simulacri del vivente, ma assumono uno statuto metaforico nei confronti prima del corpo umano, e poi anche della sua mente. A partire dalle riflessioni cartesiane, le macchine verranno elette a esempi da osservare per riuscire a comprendere il funzionamento dell'uomo, macchina eletta e fabbricata da Dio, ma pur sempre macchina. In questo senso, l'ambizione dell'umano si evolve, concretizzandosi nella volontà e nella presunta possibilità di



afferrare, narcisisticamente, attraverso le sue capacità riproduttrici e creatici, ciò che da sempre lo appassiona, ovvero l'essenza stessa del suo essere.

In questo capitolo, tenteremo quindi di ricostruire una storia della robotica che possa informare lo studio dell'interazione tra umani e robot. Prenderemo in considerazione i processi che hanno portato alla comparsa del robot come attestazione dell'ambizione dell'umano in diversi momenti (§1.1), concentrandoci in modo particolare sul processo di costruzione del concetto di macchina (§1.2) e i suoi sviluppi nel panorama delle scienze cognitive e delle scienze dell'artificiale nella metà del XX secolo.

### 1.1 L'ambizione dell'umano

I termini 'robot' e 'robotica' condividono, ovviamente, la stessa radice etimologica, ma, curiosamente, non lo stesso padre.

Il termine 'robot', come risaputo, appare per la prima volta nel 1920 nel dramma utopico fantascientifico *R.U.R.* (sigla di '*Rossumovi univerzální roboti*', traducibile come 'I robot universali di Rossum'), opera dello scrittore ceco Karel Čapek. La parola riprende la radice etimologica del termine slavo arcaico *robota*, ovvero 'lavoro', ma anche 'lavoro pesante, forzato' (con diversa accezione anche 'schiavo, servo della gleba'), ed è recuperata dallo scrittore, o per alcuni da suo fratello (Horáková e Kelemen 2008), per identificare una nuova forma di vita nata dalla volontà tecnica creatrice dell'essere umano<sup>4</sup>. L'opera descrive gli avvenimenti in una fabbrica sperduta in mezzo all'oceano diretta dal dottor Rossum, dove, allo scopo di alleviare uomini e donne dalle pesantezze del lavoro operaio, si inventa il modo di fabbricare repliche di umani per mezzo di macchine impastatrici, tini per il trattamento di protoplasma chimico e materiale organico. Tuttavia, la trama rivela come l'umanità non gioverà del risparmio della fatica, sprofondando in una spirale di vizi e indolenza e arrivando sull'orlo dell'estinzione a causa del drastico calo delle nascite, mentre i robot della Rossum ordiranno un piano per rovesciare l'ordine costituito e impossessarsi non solo dei mezzi di produzione, ma del mondo intero. Il racconto rinnova così un mito che affonda le sue radici nella notte dei tempi: quello della creatura che si rivolta contro il suo creatore. Nell'epilogo, dopo la tragica perdita degli ultimi sopravvissuti umani nella fabbrica e i vani sforzi dell'ultimo erede del genere umano di riscoprire la formula andata persa per la creazione, due dei robot iniziano a manifestare sentimenti umani, aprendo a uno spiraglio di speranza in cui la vita sulla Terra non

---

<sup>4</sup> Alcune fonti attestano un uso del termine 'robot' con un significato simile (collegato alla sfera semantica del lavoro e della fatica) in un periodo antecedente all'opera di Čapek. Ciò retrodaterebbe la nascita del termine intorno a 100 anni prima dell'uso fatto da Čapek (Curtoni 2012). Tuttavia, è indubbio che la diffusione del termine sia stata resa possibile dall'opera del drammaturgo ceco, anche attraverso una serie di passaggi di testimone tra scrittori di fantascienza come Isaac Asimov e Philip K. Dick di cui Čapek è il punto di partenza, e che la diffusione su scala mondiale di cui oggi gode il termine sia il risultato anche della popolarità dell'opera teatrale *R.U.R.*

scompare<sup>5</sup>. In sostanza, dai primi anni Venti del '900 in poi, il termine robot ha identificato tutte quelle macchine di natura artificiale che hanno lo scopo di riprodurre funzioni svolte dall'essere umano, aprendo così a un filone della fantascienza che raccoglie slanci scientifici, operazioni tecniche, paure congenite e dilemmi filosofici. Nell'inverno del 1920, quando l'opera andò per la prima volta in scena a Praga, nessun robot nel senso proprio del termine, o almeno per quello che intendiamo oggi, era ancora stato creato. Nessun artefatto al mondo era riconoscibile con questo nome. Molti ne sarebbero stati creati a partire da qualche decennio più tardi, dando ancora una volta una forma concreta a quel rapporto stretto che lega finzione e realtà, sogno e realizzazione.

Il termine 'robotica' ha un'origine molto meno nota, ma condivide con la parola 'robot' il mestiere del suo progenitore. Isaac Asimov, infatti, scrittore fantascientifico statunitense di origine russe, non ha mai fatto segreto di essere stato il primo al mondo a denotare con il termine 'robotica' un campo del sapere scientifico dedicato alla realizzazione di riproduzioni umane su vari gradi di utilità e somiglianza, sia estetica, sia funzionale, sia cognitiva. Per la prima volta il termine appare nel racconto "Bugiardo!", pubblicato nel 1938 dalla rivista *Astounding Fiction Science* e poi inserito nella famosa raccolta *Io, robot* (Asimov 1950). Asimov parla dell'invenzione di questo termine come una risposta naturale ai problemi che si era sempre posto nella sua narrativa riguardo ai rapporti tra automi e i loro creatori, ricostruendo a ritroso il nome della disciplina a partire da quello che gli scienziati avrebbero dovuto fare per creare dei robot.

Naturalmente non mi resi conto che stavo inventando una parola. La mia innocenza giovanile mi fece pensare che quella era la parola adatta, e non pensai minimamente che quel termine non era mai stato usato da nessun altro prima di me. (Asimov 1990: 15)

Nasce così nell'ambito della fantascienza una disciplina dedicata a realizzare il futuro della tecnica, un'alchimia moderna, almeno un paio di decenni prima che in alcun laboratorio scientifico o fabbrica del mondo si avessero gli strumenti per poter immaginare di costruire qualcosa di simile nelle sue fattezze e nelle sue capacità a un essere umano.

Rispetto a Čapek, nei racconti di Asimov non necessariamente le creature artificiali ordiscono trame vendicative o rivoltose nei confronti dell'umanità. Il "complesso di Frankenstein", come Asimov chiama il tema sviluppato nella fantascienza a lui contemporanea riguardo la paura del creatore rispetto alla natura malvagia e sovvertitrice delle proprie creature intelligenti, è regolato dall'invenzione del vincolo delle Tre leggi della robotica. Grazie a questo espediente, che negli anni ha riscosso un successo inimmaginato dallo stesso autore (Asimov 1990: 15), la narrativa di Asimov

---

<sup>5</sup> Alquist, l'ultimo uomo sopravvissuto all'apocalisse čapekiana, chiude la pièce con un monologo di speranza:

«E non servirà a nulla ciò che abbiamo fatto e costruito, non serviranno le città e le fabbriche, non servirà la nostra arte, non serviranno le nostre idee, eppure non perirà! Solo noi siamo morti. [...] solo tu, amore, fiorirai sulle rovine e spargerai al vento il seme della vita.» (Čapek 1921: 162)

apre per la prima volta a una visione positiva del rapporto fra robot e umani, inserendo una dimensione etica esplicita che salvaguarda l'umano e lo rimette al centro del proprio mondo<sup>6</sup> (Asimov e Frenkel 1985). In questa cornice, Asimov può quindi sviluppare riflessioni rispetto al rapporto amichevole che l'umanità può avere con le sue copie sbiadite, al ruolo che queste possono assumere per farci capire la nostra specificità in quanto umani, alle possibili riflessioni a cui le differenze rintracciate conducono. Nei racconti del *Ciclo dei Robot* non mancano ovviamente momenti di tensione, ma il rapporto instaurato da Asimov si distende su una dimensione diversa, più paritetica e solidale, che permette allo scrittore anche di immaginare tecniche evolute per la realizzazione di questi robot senza inserirle necessariamente all'interno di una cornice distopica o orrorifica.

È il caso, ad esempio, dell'invenzione letteraria del cervello positronico, il meccanismo tecnologico attraverso cui è stato possibile per gli esseri umani realizzare artefatti robotici così avanzati da mettere in dubbio il primato dell'umano nel campo dell'intelligenza, e che permette ai robot di possedere anche emozioni, comprensione evoluta dell'altro ed empatia. Naturalmente, come nel caso dei robot di Čapek, nessun corrispettivo del cervello positronico esisteva – né esiste tutt'ora – ai tempi della sua invenzione letteraria, avvenuta nel 1939 nel racconto “Robbie” (1940) ma ha contribuito in modo significativo a costruire e a far circolare nell'immaginario collettivo una visione del robot non più come minaccia all'umanità, ma come potenziale compagno. Grazie all'impiego della tecnologia, gli umani della fantascienza di Asimov erano riusciti a costruire qualcosa di estremamente efficiente, un complice intelligente che potesse accompagnarli nello sviluppo della loro storia, che potesse essere all'altezza di vivere insieme le sue avventure ma, allo stesso tempo, risultasse controllabile. Così, Asimov ha popolato interi mondi fantascientifici, in cui cervelli positronici, robot empatici e androidi convivono con gli umani, accompagnandoli in un percorso evolutivo che ora coinvolge forme di vita organica e sintetica e contribuendo ad aprire questioni filosofiche che, nella realtà, sarebbero state elaborate dalla filosofia della tecnologia e delle scienze cognitive solamente decenni più tardi.

La differenza principale, quindi, che divide l'immaginario di Čapek e Asimov è rintracciabile in una diversa rappresentazione del rapporto tra robot e umani. Mentre Čapek rinnova il tema letterario della “minaccia della creatura” instaurando una dinamica da subito conflittuale tra le parti, Asimov apre uno scenario collaborativo in cui l'interazione si fa quotidiana e le possibilità di indagine del robot si accompagnano a una nuova visione dell'umano, preso per differenza o per somiglianza alle sue creature. Ciò che ci preme evidenziare è allora proprio il ruolo fondamentale della fantascienza nella ripresa mitologica e nel rilancio filosofico e tecnico dell'ambizione dell'umano. Questo genere

---

<sup>6</sup> «[I] was tiring of robots that were either unrealistically wicked or unrealistically noble, [and] began to write science-fiction tales in which robots were viewed merely as machines, built, as all machines are, with an attempt at adequate safeguards» (Asimov e Frenkel 1985: 13).

letterario raccoglie lo slancio del mito della replica, le cui radici «si perdono nella notte dei tempi» (Asimov 1990: 15). È proprio a partire da Čapek, e sulle sue tracce, che la letteratura su questo tema e sul possibile rapporto tra umani e automi si moltiplica, rendendo la fantascienza uno dei generi più apprezzati del '900 e coinvolgendo altre forme di produzione artistica, come il cinema<sup>7</sup>. Il timore di essere sopraffatti, sostituiti o scalzati nelle nostre creature ha trovato qui terreno fertile, delineando uno scenario che ancora popola il nostro immaginario ma che si rivela catartico se accoppiato al controllo del progresso tecnologico, come nelle opere di Asimov. Ciò che di fantasia c'è nella fantascienza nutre l'immaginario filosofico che ha contribuito a far emergere nuove domande, nuove indagini e nuovi campi di ricerca.

Non è un caso che la fantascienza come genere letterario nasca a cavallo tra XIX e XX secolo, con antecedenti come le opere di Mary Shelley, H. G. Wells, Jules Verne, e che raggiunga la sua epoca d'oro intorno agli anni '30. In quegli anni, infatti, una rivoluzione partita dalle scienze fisiche e matematiche, che fino alla fine del secolo precedente erano considerate in linea di massima compiute, stava mettendo in discussione l'intero impianto delle certezze scientifiche. Nello specifico, la fisica era stata destabilizzata dall'introduzione della teoria quantistica e della teoria della relatività, che mettevano in discussione leggi che fino ad allora erano sembrate descrivere al meglio la maggior parte dei fenomeni fisici – come la meccanica di Newton e le leggi di Maxwell per l'elettricità (Crombie 1961; Geymonat 1976; Morowitz 1980); la matematica aveva appena scoperto – grazie a Gödel e Turing – di non poter essere allo stesso tempo coerente e completa come invece Hilbert aveva ipotizzato, dato che portava in seno la serpe dell'indeterminazione (Casari 1973; Odifreddi 2000; Capra & Luisi 2016). La filosofia, dunque, si insinua di nuovo al cuore della scienza, portando con sé una serie di riflessioni che avrebbero alimentato un dibattito vivo ancora oggi, nello specifico sul ruolo che l'osservatore assume nella comprensione del processo indagato (Schrödinger 1944; Heisenberg 1958; von Foerster et al. 1974; Morowitz 1980; Bocchi & Ceruti 1985; Ceruti 1989; Damiano 2010; Capra & Luisi 2016).

La fantascienza emerge così in un momento in cui la frattura delle scienze dure apre uno spiraglio per l'immaginazione del futuro non più ingabbiato dal determinismo scientifico, e raccoglie la sfida antica di immaginare come riprodurre in qualche modo l'essere umano. Si apre quindi un panorama inedito in cui lo stato della tecnologia dell'epoca, le idee alla base degli account scientifici e filosofici

---

<sup>7</sup> Il primo film che rappresenta un robot è *Metropolis* (1927), film muto del regista Fritz Lang considerato l'antecedente dei film di fantascienza del secondo Novecento. Da quel momento in poi, complice anche la grande ascesa del cinema, i prodotti cinematografici sui robot si sono moltiplicati, arrivando a costituire un filone indipendente della fantascienza. Tra i più celebri in Occidente, dichiaratamente ispirati a *Metropolis*, si ricorda la saga di *Guerre stellari* (1980-2018) di George Lucas, *Blade Runner* (1982) di Ridley Scott e *Terminator* (1984) di James Cameron. Per una panoramica delle possibili influenze del cinema di fantascienza sulla robotica: Telotte (2016), Saffari et al. (2021); per le rappresentazioni del robot nel panorama mediale contemporaneo: Schonfield & LeRoy (2018), Locatelli & Toniolo (eds. 2022).

e le proposte di un immaginario possibile hanno fiorito nella mente di quelli che sarebbero stati i primi robotici della storia. Ad esempio, George Devol e Joseph Engelberger, considerati gli ideatori del primo braccio robotico digitalmente operato e programmabile della storia, hanno dichiarato di aver avuto l'idea dell'UNIMATE proprio leggendo i racconti di Asimov<sup>8</sup> (Stone 2018); Rodney Brooks, robotico celebre per aver inaugurato il paradigma della *Nouvelle AI*, racconta di essere rimasto stregato da bambino dalle avventure di HAL 9000<sup>9</sup>, il robot di *2001: Odissea nello spazio* (1968), e per questo aver intrapreso la sua carriera (Brooks 1990, 2002).

Nella frattura del panorama scientifico e filosofico del '900, dunque, la fantascienza si infila con la sua doppia anima, letteraria e filosofica, a rinforzare un connubio tra fantasia e possibilità tecnica che viene sapientemente raccolto dall'ingegneria robotica, e che mette in scena con variazione la sempiterna ambizione dell'umano, ovvero l'intento di replicare la natura umana attraverso l'impiego della tecnica. Al fine di muovere i primi passi nel percorso di indagine sull'interazione tra umani e robot, si ritiene cruciale comprendere il cammino di evoluzioni e cambiamenti che hanno portato all'attuale possibilità di realizzare robot capaci di entrare in relazione con gli esseri umani. Si comincerà quindi con il mostrare come, se di robot si inizia a parlare solamente all'inizio del '900, il *concetto* di robot sia molto più antico, «forse quanto la fantasia umana» (Asimov 1990: 15).

### 1.1.1 Il mito della replicabilità

Si può affermare senza esitazione, come ha sottolineato Asimov, che il robot in quanto macchina che si muove ha una storia antica e che il concetto su cui si poggia ha affrontato una lunga evoluzione, contaminandosi con le filosofie e sfruttando le tecniche che hanno caratterizzato i diversi periodi nella storia dell'umanità. Ciò che è necessario capire è come questa lunga storia possa informarci su quello che è il robot oggi.

Infatti, se con robot intendiamo un insieme di meccanismi elettronici che, sapientemente collegati e raccolti da involucri bio-ispirati, mettono in scena un comportamento simil-umano, e che allo stesso tempo garantiscono un grado di autonomia nel comportamento stesso, allora certo, non possiamo che rivolgere la nostra attenzione alle invenzioni della seconda metà del XX secolo. Ma come si è potuti arrivare alla loro costruzione? L'attività odierna di realizzazione di robot capaci di interagire con gli

---

<sup>8</sup> «Over and over, the advice was, “Don’t call it a robot. Call it a programmable manipulator. Call it a production terminal or a universal transfer device.” The word is *robot* and it *should* be *robot*. I was building a robot, damn it, and I wasn’t going to have any fun, in Asimov’s terms, unless it was a *robot*. So I stuck to my guns» (Joseph Engelberger, citato in Asimov e Frenkel 1985: 25)

<sup>9</sup> HAL 9000 è stato realizzato su consulenza di un gruppo di informatici, robotici e scienziati cognitivi, tra cui Marvin Minsky (Kirby 2011). Curiosamente, Brooks inizierà la sua carriera e darà vita al suo modello alternativo per la robotica situata proprio a partire da una critica ad alcune idee di Minsky sulla natura computazionale della cognizione (Brooks 1991a, 1991b).

esseri umani non è che l'ultima fase di un processo di immaginazione della cui origine non si ha memoria pratica, la cui evoluzione è tutt'altro che scontata.

Nonostante alcuni abbiano cercato di fornire interpretazioni della storia del concetto di robot che ponesse in una soluzione di continuità storica i robot costruiti a partire dalla seconda metà del XX secolo e le invenzioni reali o immaginarie che si sono succedute nei secoli (Meyer 2012; Wood 2002), questi tentativi si rivelano utili più a una storia dell'immaginario relativo alle macchine semoventi che all'obiettivo di comprendere come si sia arrivati a realizzare i robot novecenteschi. Ad alcuni (Losano 1990; Sini 2009; Israel 2004; 2015) va riconosciuto il merito di aver costruito un percorso in cui l'invenzione del robot si configura come sorretta da una serie di elaborazioni e traguardi – tecnologici e filosofici – senza i quali sarebbe stato impossibile arrivare a realizzare il robot degli anni '70. Tuttavia, se puntiamo l'attenzione su ciò che rende i robot di oggi simili ai loro antenati, si rischia di perdere di vista l'importanza della progressiva evoluzione della figura robot, fatta di scarti e differenze che sottolineano sia l'avanzamento scientifico, riguardante il funzionamento delle macchine, sia quello della conoscenza, che all'avanzamento si accoppia nel tentativo filosofico di guardare le creazioni per comprendere qualcosa circa la natura dell'essere umano. Se Pigmalione, protagonista del celebre mito, avesse incontrato SHAKEY – il primo robot autonomo costruito nella storia (§1.3.4) – al posto di Galatea, probabilmente sarebbe scappato a gambe levate invece che innamorarsene; i meccanismi che permettono a iCub (Metta et al. 2010), robot autonomo dalle sembianze infantili, di muoversi non sono minimamente paragonabili a quelli che animavano le porte automatiche della città di Alessandria, anche se possono essere considerati un loro antichissimo antenato; cosa avrebbero detto le genti del Medioevo se invece che osservare le teste parlanti di Lullo si fossero trovati davanti Kismet (Breazeal 2000)? In sostanza, se da una parte è utile non fissare confini troppo rigidi per ricostruire una storia della robotica, dall'altra è importante non renderli troppo permeabili fino ad includere ogni marchingegno automatico dalla la ruota in avanti. Partendo, quindi, dall'oggetto di analisi della nostra trattazione – ovvero, il robot umanoide contemporaneo – cercheremo di astrarne alcune caratteristiche e cercare a ritroso come queste siano arrivate ad essere immaginate o realizzate nel corso della storia. In questa ricostruzione segnata da ricorrenti discontinuità, ci interessa vedere come l'ambizione di realizzare copie e doppi da parte degli esseri umani si sia continuamente intrecciata con la volontà di comprendere il funzionamento del corpo e della mente umana attraverso le macchine. Perciò, ciò che pertinentizzeremo come oggetto di indagine è il tentativo di riprodurre l'umano sotto qualche aspetto, che sia questo comportamentale, estetico o cognitivo.

Dobbiamo quindi affidarci a un processo di astrazione preliminare della figura del robot novecentesco, in cui rintracciare le invarianti che permettono di identificare il robot come ultimo

anello di una lunghissima catena di realizzazioni. Tali invarianti, che ci consentono di identificare il nostro oggetto di indagine, sono individuabili in almeno tre caratteristiche: 1) replicazione del movimento, o dell'illusione del movimento autonomo; 2) natura sintetica o artificiale della creatura; 3) somiglianza a forme di vita biologiche, animali o umane. Per questo, dalla nostra indagine vengono escluse quelle forme di riproduzione meccanica che non cercano di imitare il comportamento, il movimento, o le sembianze del corpo umano o animale. Ad esempio, per quanto oggetto di un'approfondita storia dell'automazione e filosofia della tecnica, non rientrano nei nostri studi i meccanismi di aperture automatiche di porte rintracciabili nelle descrizioni dei palazzi alessandrini del III secolo a.C. (Rossi e Russo 2009, §21.1.1). Per lo stesso motivo, includiamo d'altra parte nel nostro corpus anche esempi provenienti dalla mitologia e dalla letteratura, che non avrebbero mai potuto essere realizzati data la loro comune origine magica o divina, ma che hanno contribuito in modo sostanziale a nutrire l'immaginario della vita artificiale. Ad esempio, come verrà sottolineato in seguito e come riportato in diverse ricostruzioni della storia della robotica (Losano 1990; Sini 2009; Israel 2004; Mayor 2018), il mito del Golem assume un ruolo fondamentale nella modellazione dello scenario antecedente ai robot propriamente detti, e contribuisce a sviluppare quel topos della "creatura ribelle" che arriverà prima a Shelley e poi Čapek, approdando quindi direttamente nella scena Novecentesca.

### *1.1.2 Statue, miti e teste parlanti*

Le prime attestazioni di costruzione di macchine con intento di replicazione di alcuni aspetti e funzioni dell'umano sono rintracciabili nel periodo ellenistico. Infatti, anche se i meccanismi implementati in queste macchine fossero già conosciuti ben prima, i primi esempi del loro impiego per realizzare macchine semoventi dalla foggia umana si ritrovano nell'opera dello scienziato Erone di Alessandria (Rossi e Russo 2009), vissuto tra il III e il II secolo prima di Cristo e passato alla storia come il primo costruttore di macchine umanoidi. La parola designata per denotare questo particolare tipo di macchine nasce proprio in questa congerie: *autómata* (αὐτόματα) è «ciò che si muove da sé, ma anche ciò che *pensa* da sé, se risaliamo al termine sanscrito *mata*, pensiero, donde il greco *máino*: ho in anima (nella *mens*), bramo, desidero, agogno, tendo, e anche irrompo e cerco» (Sini 2009: 9-10). L'ambizione dell'umano si concretizza quindi per la prima volta in questa sede, in marchingegni semoventi che danno l'impressione di voler replicare ciò che anima l'umano dal di dentro attraverso meccanismi automatici.

In genere, i meccanismi utilizzati per far muovere le macchine o per far loro espletare una qualche funzione che mimasse il comportamento umano provenivano dallo sfruttamento delle forze naturali, come l'acqua, il vento o sistemi di pesi. Nonostante i meccanismi fossero semplici, l'effetto suscitato

in chi si trovasse alla presenza di queste statue è narrato come meraviglioso e divertente, motivo principale per cui marchingegni simili venivano costruiti. Figure come Ctesibio (III sec. a.C.) e Filone di Bisanzio (III sec. a.C.) sono passate alla storia come importanti antecedenti per l'ingegneria e l'architettura successiva, ma anche come impareggiabili costruttori di meccanismi mimetici e statue dalle fattezze umane. Da una disamina degli esempi raccolti dalla letteratura e arrivati fino a noi, si può parlare di tre tipologie di marchingegni: «macchine laboriose, macchine divertenti e macchine da guerra» (Sorge 2019: 82). Alle prime due tipologie possono essere ricondotti numerosi artefatti dalle sembianze umane, che mostravano diverse capacità a seconda delle funzioni che erano destinati a svolgere. Del tipo delle “macchine laboriose” è l'*Automate Therapaenis*, che riproduceva le sembianze di una fanciulla e svolgeva le funzioni di una cameriera ai banchetti dei nobili. La statua semovente era dotata di un meccanismo che trasferiva vino e acqua da due recipienti posizionati all'altezza del busto alla brocca che teneva in mano, sfruttando dei tubi che passavano lungo un braccio. Un'altra serie di cannule attraversava l'altro braccio, rendendole possibile il gesto di versare i liquidi: quando l'ospite appoggiava una coppa sulla mano libera, il peso di questa faceva abbassare la mano e attivava la mescita di vino e, in un secondo momento, dell'acqua, fermandosi quando il recipiente veniva sollevato. In più, all'altezza delle gambe era nascosto un motore meccanico, attivabile dalla caduta ritardata di un sacchetto di fagioli secchi, che azionava le ruote e consentiva alla statua di muoversi in linea retta da un punto all'altro della stanza (Henderson 2009). Una odierna riproduzione si può ammirare al Museo della Tecnologia Greca Antica, a Katakolo Ileias, nel Peloponneso.

Tra le “macchine divertenti” può senza dubbio rientrare la statua di Memnone per i greci, o Amenofis per gli egizi, eretta a Tebe, che quando colpita dai raggi del sole all'alba emetteva suoni simili a quelli prodotti da una lira (Cohen 1966: 8). A lungo si è speculato sul meccanismo che rendeva possibile che la statua cantasse al sorgere del sole, come ad esempio un sistema di linguette vibranti che colpite venivano riscaldate producendo il suono. Per molti interpreti, però, si trattava più di un trucco, come per Athanasius Kircher che ipotizzò che un reale arpicordo fosse nascosto all'interno della statua e che venisse suonato all'occorrenza. Per altre statue che mostravano la stessa capacità canora, come le statue di Alessandria o quella di Esculapio, era stato dimostrato che altri stratagemmi erano stati trovati dagli ingegnosi costruttori, come sistemi di tubi che facevano viaggiare la voce da una stanza attigua fino alla bocca della statua, nascondendo quindi un intervento umano (Ibid.: 17).

Per questi esempi realmente realizzati nell'antichità c'è da dire che, nonostante l'impiego e la costruzione non provenissero da, né nutrissero, alcuna pratica scientifica, la loro realizzazione conseguiva una attenta osservazione della natura dell'umano e dei fenomeni naturali in generale. Questi meccanismi, infatti, incarnavano i principi di una scienza rudimentale, che ancora non era stata



sviluppata sistematicamente ma che per la prima volta metteva in forma l'ambizione dell'umano di avvicinarsi agli dèi che popolavano le loro mitologie.

A tal proposito, è impossibile non parlare del ruolo che la mitologia occupa per lo sviluppo dell'ambizione dell'umano di creare proprie copie e doppi. È nella mitologia, d'altronde, che per prima si realizza la fantasia di riuscire nell'impresa. Mancando delle tecniche necessarie per creare ciò che nella fantasia era già vivido, infatti, il ricorso alle narrazioni mitologiche, agli strumenti o poteri divini, a formule magiche o incantesimi e sotterfugi funziona da primo tramite per mettere in scena ciò che l'immaginazione aveva formulato: «tra la fantasia e l'esatta conoscenza, tra la pratica e la tecnologia, vi è uno stadio intermedio: quello della magia» (Mumford 1961: 53; *infra* §1.2).

Per queste ragioni, dunque, risulta naturale sottolineare il ruolo primario che viene assunto dalla tradizione greca, tra le prime civiltà antiche a riportare testimonianze fantastiche di automi semoventi nei racconti mitologici, nel delineare il rapporto tra immaginario e realizzazione. In particolare, la prima attestazione di creature con vita e movimenti artificiali è data nell'*Iliade*, in cui il dio zoppo Efesto è affiancato da due ancelle d'oro da lui stesso costruite mentre realizza dei tripodi che potevano andare da un posto all'altro nella sala del banchetto riservata agli dèi:

[...] avea per mano  
Dieci tripodi e dieci ad ornamento  
Di palagio regal. Sopposte a tutti  
D'oro avea le rotelle, onde ne gioisse  
Da sé ciascuno all'assemblea de' numi,  
ed a sé ne tornasse onde si tolse:  
maraviglia a vedersi! [...]  
(*Iliade*, Libro XVIII, vv.509-15)

Efesto, dio della metallurgia e campione di inventiva, ha la capacità divina di forgiare strumenti e marchingegni di eccelsa fattura. Questa viene messa al servizio di tutti gli altri dèi dell'Olimpo, che vengono dotati di armi e dispositivi divini come l'arco e le frecce d'oro di Apollo o le armature infrangibili di Achille o Memnone. Ma ciò che è di maggiore interesse per noi è che la sua abilità si realizza anche nella creazione di altre figure mitologiche. È il caso, ad esempio, di Pandora e del suo vaso, o del gigante di bronzo a guardia della città di Creta Talos (cfr. Mayor 2018, cap. 7). La figura di Efesto stabilisce un precedente per la figura di creatore di artefatti semoventi con sembianze umane non solo nella cultura greca, e quindi per la cultura mediterranea in generale, con figure che

ripropongono lo stesso personaggio nella figura latina con Vulcano, o nella cultura araba, ma anche nella mitologia norrena (Cohen 1966).

Di pari diffusione e rilevanza è la sua controparte umana nella mitologia greca, cioè Dedalo. La natura umana di questo personaggio contribuisce a rendere ancora più definito il mito dell'uomo capace di realizzare oggetti e strumenti di straordinaria ingegnosit  e creativit . Leggendaria artefice e inventore, a Dedalo sono attribuite numerose creazioni, come gli strumenti della falegnameria, e opere di ingegno, tra cui l'ariete di bronzo di Siracusa, il toro realizzato per la moglie di Minosse, o il labirinto costruito per quest'ultimo, in cui poi rimarr  intrappolato con suo figlio Icaro e da cui fuggir  assemblando un paio di famosissime ali di cera (Mayor 2018).

Altra figura centrale nella mitologia greca per il suo ruolo di creatore di artefatti con fattezze umane   il personaggio di Pigmalione. A differenza della natura specifica e strumentale di entrambi i personaggi precedenti, per , la figura di Pigmalione ci consente di introdurre un ulteriore ambito in cui l'ambizione dell'umano si concretizza, ovvero quello dell'arte. Il mito racconta di come Pigmalione, scultore cretese, si fosse innamorato della sua stessa opera, una statua eburnea raffigurante un nudo femminile straordinariamente simile a una fanciulla in carne ed ossa. Tanto era il desiderio di poter amare la statua in modo umano che la dea a cui Pigmalione rivolge le sue preghiere, Afrodite, sceglie di accontentarlo e donare la vita al pezzo di marmo<sup>10</sup>.

Sebbene questa breve carrellata possa gi  far rendere conto che l'intento di realizzare oggetti e manufatti con fattezze umane fosse un tema ricorrente nella cultura greca classica, resta da chiarire il modo in cui questa ambizione viene concettualizzata e perduri cos  nel tempo. Il rapporto tra immaginario e realizzazione pratica nella Grecia ellenica   pi  problematico di quanto non possa sembrare. Infatti, nonostante il nutrito immaginario di esempi di realizzazioni di opere d'ingegno cos  simili all'umano, sul versante pratico le realizzazioni non erano accompagnate da una messa in forma della tecnica. In altre parole, le potenzialit  fantastiche che hanno contribuito a forgiare i miti che ancora oggi popolano il nostro immaginario non hanno avuto seguito nel tentativo di essere realizzate in artefatti reali, se non nell'intento di meravigliare e imitare qualche caratteristica umana superficiale. La distanza che era posta tra fantasia e realizzazione era di molto maggiore rispetto a quella che possiamo misurare oggi tra le nostre idee e le possibilit  che abbiamo di realizzarle. Nel mondo greco il ruolo dei personaggi mitologici che costruiscono artefatti con sembianze umane   raccolto pi  sotto il rispetto della creazione di un'etica, piuttosto che di una tecnica. Non era centrale la possibilit  di realizzare ci  che si immaginava, quanto pi  il potere direttivo che tali parabole avevano per la condotta dell'umano e per guidare il suo atteggiamento nei confronti della natura in generale. Esempio in questo senso   il mito di Prometeo, caro non solo ai greci ma ripreso e

---

<sup>10</sup> Per una trattazione approfondita delle figure mitologiche greche nella storia della robotica, si veda Mayor (2018).

rielaborato nel corso del tempo con l'intento di ribadire, o ridefinire, il confine tra uomo e divino<sup>11</sup>. Come noto, il mito racconta del titano Prometeo che, dopo aver creato lui stesso l'uomo dalla terra<sup>12</sup>, rubò dal carro del Sole il fuoco sacro degli dèi per donarlo agli uomini. Ciò scatenò l'ira di Zeus, che incatenò Prometeo per sempre al fianco di una montagna mentre un'aquila gli rodeva il fegato che ogni notte si rigenerava. Il mito è interessante non solo perché funge da monito per chi tenta di eguagliare le imprese divine o di riprodurre i prodigi, ma anche per le figure mitologiche a questo legate. Primo fa tutti la figura di Pandora, prima donna sulla Terra, creata da Efesto e ingannata da Zeus che nel vaso donatole ripone tutte le disgrazie che avrebbero afflitto il genere umano di lì in avanti: pazzia, gelosia, vecchiaia, malattia e vizio. Solo la Speranza resta sul fondo del vaso e, insieme alla Curiosità donata a Pandora da Ermes, resterà a disposizione degli uomini.

Il mito di Prometeo acquista allora una primaria importanza, perché raccoglie ed esemplifica la disposizione classica nei confronti dell'impresa della conoscenza e della scoperta. Ciò che principalmente resta e si diffonde del mito di Prometeo è il peccato di *hybris* commesso dal titano: «l'irriverenza sacrilega, il ricorso all'inganno e al sotterfugio, la volontà di andare oltre i limiti concessi dalle divinità» (Battistini 2015: 192) sono ciò che lo condanna alla sua pena infinita e che prevale nella sua lettura, almeno fino a un certo momento. Prometeo è incatenato alla montagna, punito per aver rivelato l'armamentario della tecnica agli uomini, che saranno aiutati nella loro impresa di costruzione della conoscenza da quegli strumenti che erano prerogativa divina prima del dono e che nutriranno quella Curiosità cara a Pandora che si unisce alla Speranza rimasta sul fondo del vaso. Speranza, soprattutto, di non valicare quel limite inconoscibile tra umano e divino e quindi di non essere puniti come Prometeo. Questa visione del limite sarà di fatto l'interpretazione del mito più diffusa in epoca ellenistica prima e romana poi, arrivando quindi al Medioevo mescolata a influenze cristiane con tutto il suo portato cautelativo nei confronti della scoperta del funzionamento del creato<sup>13</sup>.

D'altra parte, a partire da un certo momento storico e in concomitanza con alcune riflessioni filosofiche, il mito di Prometeo sarà interpretato anche sotto un'altra luce, che mette in risalto non la punizione divina al gesto del dono ma il perseguimento dell'opera di conoscenza stessa reso possibile agli uomini da Prometeo. In questo senso, un ruolo centrale sarà assunto da altre caratteristiche: la

---

<sup>11</sup> Le riprese del mito di Prometeo sono disseminate per tutta la storia della filosofia occidentale, oltre che in letteratura, cinema e raffigurazioni artistiche, rendendo impossibile raccogliere qui le occorrenze delle sue apparizioni. Si tornerà sul mito per darne una lettura moderna in §1.2.1. Per una rassegna almeno parziale nella storia della filosofia, si veda Battistini (2015); Rossi (1962:177-188).

<sup>12</sup> La creazione dell'uomo da parte di Prometeo è presente nella versione narrata da Ovidio, in cui il titano modella la sua creatura mescolando terra e acqua e aggiungendo il calore del fuoco fino a formare l'uomo «in effigiem moderantium cuncta deorum» (Ov, *Met.*, I, 83), ovvero a immagine e somiglianza degli dèi (Battistini 2015: 192).

<sup>13</sup> La condanna a chi travalica il confine posto dal principio regolativo del “ne quid nimis”, ovvero “niente di eccessivo”, è comune a entrambe le culture, greca e latina, e si protrae come monito cristiano nella filosofia naturale almeno per tutto il Medioevo (Battistini 2015).

curiosità della scoperta, l'astuzia, l'intraprendenza e la generosità di Prometeo, che si sacrifica per donare agli uomini non solo il fuoco, ma anche «l'insieme di ogni possibile *technè*» (Ibid). Per arrivare a prediligere la lettura ottimista del mito, però, sarà fondamentale una rivoluzione nella concezione del rapporto tra umano e natura, che non arriverà prima del XVII secolo.

È nella tenuta di due miti, quindi, che si può rintracciare la tensione che guida lo sviluppo dell'ambizione dell'umano nella mentalità greca: da una parte quella tendenza a duplicarsi, il mito della replicazione, testimoniata sia dalla mitologia divina che dalle figure umane esaltate come ingegnosi creatori; dall'altra il monito di Prometeo, che minaccia l'ambizione ponendole un freno.

Tale concezione duplice e mitologica si protrae fino al Medioevo. Infatti, nel Medioevo, la stessa componente magica e fantastica che faceva da sfondo alle creazioni greche continua a fornire l'appiglio per l'ambizione dell'umano di replicarsi, sia in ambito tecnico, sia letterario. In ambito tecnologico, in quel periodo storico si era arrivati a sviluppare ruote dentate, camme, mulini a vento e ruote idrauliche, ma soprattutto mantici pneumatici, importanti per la realizzazione degli automi, che furono utilizzati per la prima volta per costruire organi a canne (Heppenheimer 1985: 30; Losano 1990: 50-62).

È, tuttavia, la componente letteraria che ancora ricopre il ruolo primario. Sono arrivate fino a noi numerose leggende di personaggi medievali ambigui che, alla loro già risaputa indole curiosa, aggiungevano un tocco di mistero e stramberia. In molti casi, le leggende riportano di oggetti simili a quelli che abbiamo esposto fin qui, ovvero creature infuse di vita artificiale, statue semoventi, o teste parlanti. È il caso, ad esempio, di Gerberto di Aurillac, eletto papa Silvestro II nel 999 d.C., il quale era creduto essere un mago dato il suo interesse per le arti occulte, la matematica e la negromanzia, ma anche per la sua profonda conoscenza della cultura araba (Graf 1996). A nutrire questa immagine era anche il fatto che si narrasse che Silvestro II fosse in possesso di una testa parlante, che consultava ogni qual volta dovesse prendere una decisione (Cohen 1966: 19-29). Di un'altra testa parlante si narra fosse costruttore Ruggero Bacone (1214-1292), il *Doctor mirabilis* che profetizzava macchine volanti e carri che si sarebbero mossi da soli<sup>14</sup>. Tra i nomi illustri da aggiungere alla lista dei possessori di teste parlanti, Alberto Magno (1204-1282) spicca come il protagonista della leggenda più curiosa. Si narra che possedesse un «robot [che] sapeva rispondere alle domande e risolvere problemi, e si azzardò perfino a rivolgere il saluto a un grande allievo del suo padrone, il futuro san Tommaso d'Aquino, il quale, convinto che il robot fosse una creatura del demonio, non esitò a consegnarlo alle fiamme» (Cohen 1966: 22).

---

<sup>14</sup> «Se la natura, poi, è aiutata dal sapere e dall'ingegnosità umani, allora i risultati saranno quasi incredibili per gli inesperti. Si posso no costruire navi che navigano senza l'opera dei rematori, macchine per volare nei cieli o per esplorare gli abissi del mare controllate da un solo uomo, congegni adatti per sollevare pesi enormi, strumenti per condizionare la volontà altrui». (Bacone, *De secretis operibus artis et naturae*, IV)

Parallelamente alla serie di leggende sulle teste parlanti di possesso di personaggi illustri dell'epoca medievale, la storia dell'ambizione dell'umano di replicare alcune funzioni che gli appartenevano in esclusiva acquista una nuova formulazione. Vero precursore di questa nuova corrente è il pensiero di Raimondo Lullo (1235-1315), il *Doctor illuminatus* celebre per aver elaborato l'*Ars magna* (*Ars compendiosa inveniendi veritatem seu ars magna et maior*, Lullo 1274), un metodo per collegare i concetti fondamentali tramite simboli allo scopo di raggiungere la verità, una sorta di logica meccanica che trova secondo letteratura una realizzazione empirica nella “macchina per pensare” (Borges 1937, Eco 1993). L'intento primario era quello di elaborare un modo di comunicazione universale che consentisse di convertire gli infedeli al cristianesimo grazie al dispiegamento delle connessioni logiche e ontologiche del mondo, ma ciò che risulta di maggiore interesse in questa sede è che, con la macchina, Lullo cercasse di condensare gli elementi di una conoscenza universale, raccogliendo i principi delle singole scienze e combinandoli per mezzo di cerchi concentrici. Il presupposto consisteva nel pensare che la ragione fosse slegata dalla fede e che tutto lo scibile fosse costituito da un numero finito di concetti e categorie di per sé evidenti. Combinandoli, Lullo credeva si potesse arrivare alla verità attraverso la macchina, senza sforzarsi di ragionare o indagare (Cohen 1966: 25-27). Nella “macchina per pensare” di Lullo è racchiusa la prima attestazione di un tentativo di astrazione del pensiero logico al fine di riprodurlo tramite strumenti costruiti dall'uomo, e ciò costituirà un importante antecedente ripreso ed elaborato da pensatori come Giordano Bruno, principale testimone dell'ars lulliana e traghettatore del suo pensiero ai posteri, o Leibniz, approfondito conoscitore dell'*Ars magna* e dell'ars combinatoria a questa legata.

Ad alcuni, tuttavia, non bastava cercare di replicare l'umano attraverso espedienti meccanici e tecniche di riproduzione del pensiero logico. È in questo quadro che si inserisce l'alchimia, disciplina antenata della chimica odierna, come scienza della riproduzione dell'uomo da parte dell'uomo senza l'impiego di tecniche meccaniche, ma attraverso l'indagine dei legami oscuri della materia. Ad accomunare i diversi tentativi di produrre una copia dell'umano è senza dubbio l'aura magica: ad esempio, il processo di costruzione delle teste o degli automi non è mai svelato, ricorrendo a stratagemmi occulti e a tecniche di creazione sconosciute che fanno spesso riferimento al processo alchemico. Quanto al legame di Lullo con la magia, questo è più che noto e spesso arriva a oscurare l'importanza del ruolo che la filosofia lulliana ha ricoperto nello sviluppo del pensiero filosofico medievale (Eco 1993; Cohen 1966).

Insieme all'ingegneria, alla letteratura e all'arte, l'alchimia come disciplina che incarna lo studio del principio magico della natura è quindi un altro campo fondamentale da prendere in considerazione per ricostruire l'archeologia dell'ambizione dell'umano di ricreare suoi doppi. Con i progressi compiuti dall'alchimia, nel primo Cinquecento ci fu chi affermò di poter creare non solo teste di

ottone, ma anche interi esseri umani in miniatura, detti *homuncoli*<sup>15</sup>. Addirittura, l'alchimista svizzero Paracelso, forse il più popolare tra gli scienziati del suo tempo, descrive passo per passo le istruzioni per la sua creazione<sup>16</sup>:

Chiudete per quaranta giorni in un alambicco del liquido spermatico di un uomo; e che si putrefaccia fino a che cominci a vivere e a muoversi, ciò che è facile constatare. Dopo questo tempo, apparirà una forma simile a quella di un uomo, ma trasparente e quasi senza sostanza. Se, dopo questo, si nutre tutti i giorni questo giovane prodotto, prudentemente e accuratamente, con sangue umano e lo si conserva durante quaranta settimane ad un calore costantemente uguale a quello del ventre di un cavallo, questo prodotto diventerà un vero e vivente fanciullo, con tutte le sue membra, come quello che è nato dalla donna, soltanto molto più piccolo. Bisogna allevarlo con molta diligenza e molte cure, fino a che egli cresca e cominci a manifestare intelligenza.

Tuttavia, Paracelso confessò di non aver mai tentato di eseguire la procedura, visto che sarebbe stato come sfidare Dio nella sua essenza, cioè quella di creatore assoluto (Prasad 2014).

Giunti fin qui, possiamo raccogliere gli esempi finora enumerati sotto l'etichetta di *macchine idraulico-pneumatiche*: i principi meccanici che animavano artificialmente gli artefatti simil-umani (e non solo) costatavano di principi elementari dovuti allo sfruttamento di alcuni forze e fenomeni fisici, come l'acqua, il vento, il peso degli oggetti e così via, messe in pratica all'occorrenza ma non ancora oggetto di una indagine specifica e metodica. L'aspetto caratteristico di questo tipo di macchine è che esse «non producono alcuna immagine precisa dell'uomo inteso come macchina, tantomeno forniscono un'idea di come le macchine possano essere intese come esseri inanimati» (Israel 2004: 63). La ragione è che, a differenza di praticamente tutta la storia successiva del pensiero umano, la società greco-romana e quella medievale non aveva sviluppato una visione meccanicistica del mondo: il passaggio dal “mondo del pressappoco” all’“universo della precisione” era ancora di là da venire (Koyré 1948).

È in questa tensione che si giustifica la mancanza dello sviluppo di una *techne* come correlato di un *episteme*, votata invece alla *praxis* come accade in questo caso. Secondo Koyré, diverse sono le cause del motivo per cui la *techne* greca non abbia saldato la sua riflessione sui modi di realizzare artefatti con una impresa generale di conoscenza, producendo di fatto una tecnologia, anche se i presupposti sembravano essere sufficienti. Il modo greco, d'altronde, aveva sviluppato un armamentario tecnico matematico e geometrico che molti secoli dopo, quasi invariato, avrebbe posto le basi per la nuova scienza. Tra le cause analizzate, di natura socioculturale, economica e politica, la più interessante in questa sede, e anche secondo Koyré, risulta essere la mancanza di una

---

<sup>15</sup> La fortuna della figura dell'*homunculus* in filosofia è ben documentata, cfr. Prasad (2014).

<sup>16</sup> Il passo è tratto dall'opera di Paracelso *De Natura Rerum*, citata da *Gli occultisti*, a cura di Giulio Alliney, Garzanti, Milano, 1951, pp. 82-83.

sistematizzazione matematica della natura nel mondo greco, rendendo così impossibile lo sviluppo di una fisica e di una tecnologia.

In effetti, fare della fisica nel *nostro* senso del termine – non nel senso che Aristotele dava a questo vocabolo – vuol dire applicare al reale nozioni rigide, esatte e precise della matematica e, in primo luogo, della geometria. Impresa paradossale, se mai ve ne furono, poiché la realtà, della vita quotidiana in mezzo alla quale viviamo e stiamo, non è matematica. E neppure matematizzabile. Essa è il dominio del movente, dell'impreciso, del «più o meno», del «pressappoco». [...] Ne risulta che voler applicare la matematica allo studio della natura è commettere un errore e un controsenso. (Koyré 1948: 90)

Non era possibile, dunque, per i Greci riuscire ad immaginare un'opera sistematica di realizzazione degli automi, né tanto meno un uso scientifico dei propri modelli. In un'epoca come quella antica in cui fioriscono gli esempi di macchine automatiche, in cui la filosofia del materialismo atomico di Democrito e Lucrezio apre a una visione scomponibile del corpo e della mente umana, l'idea di una macchina vivente non riesce ad emergere. Tale situazione resterà pressoché invariata fino alla rivoluzione scientifica e alla concomitante formulazione di un pensiero meccanicista intorno alla metà del diciassettesimo secolo. C'è da aggiungere che i pensatori del Medioevo, come abbiamo accennato, propenderanno molto di più per una riflessione basata sullo spiritualismo cristiano, contribuendo a portare avanti l'ambizione dell'umano maggiormente sul lato intellettuale che su quello delle realizzazioni empiriche (Israel 2004, 2015; Cohen 1966). Ciò che renderà possibile passare da una visione giocosa e utilitaristica della produzione di automi a una in cui si scopre la validità dei meccanismi studiati costruendo macchine è proprio il passaggio dal mondo del “pressappoco” all'universo della “precisione”, ossia lo sviluppo di uno sguardo specifico che mette il concetto di precisione, di misura e di commensurabilità al centro dell'impresa conoscitiva della scienza (Koyré 1965, 1967). Infatti, anche se il termine ‘automa’ viene ampiamente impiegato fin dalla Grecia classica per denotare realizzazioni artificiali con la capacità di “muoversi da sé”, lo sviluppo di meccanismi complessi che rendono effettivamente gli automi capaci di riprodurre più o meno accuratamente, ma in modo scientifico, alcune funzioni biologiche avverrà solamente a seguito di questa rivoluzione. L'osservazione attenta della natura non era ancora sufficiente, era necessaria la volontà di comprenderla per governarla (Hacking 2005, cfr. *infra*, §1.2). Questa volontà si formerà nei secoli con lo sviluppo della tecnica e con una serie di scoperte scientifiche, evolvendo le tecniche di realizzazione degli artefatti fino a far scattare una scintilla nella mente degli scienziati moderni.

All'analisi delle prime attestazioni di macchine che alimentano e realizzano per prime l'ambizione dell'umano, allora, nel prossimo paragrafo faremo seguire un lavoro di ricostruzione dei concetti e degli sviluppi epistemologici e filosofici che hanno creato le condizioni affinché il robot, come macchine novecentesca risultato dell'ambizione dell'umano, si potesse realizzare. Vedremo come, senza il *concetto* di macchina per come è stato elaborato a partire dal XVII secolo, il *concetto* di robot

non avrebbe potuto prendere forma per quello che osserviamo oggi. La sequela di realizzazioni di artefatti artificiali, quindi, fino alla nascita della robotica come disciplina, va letta nella sua completezza come una serie di *exempla* in un processo che, mentre si dispiega, si modifica in base alle influenze del pensiero con cui viene contaminato. È necessario dunque passare ad analizzare come il concetto di macchina si sia evoluto, tracciando un sentiero che, almeno a partire da Cartesio, ha condotto l'umanità a concepire la robotica nella metà del XX secolo.

## 1.2 Il concetto di macchina

### 1.2.1 Macchine e automi

Il panorama scientifico e filosofico che si apre alla fine del XVI secolo è forse uno dei più interessanti se visto nella prospettiva di una ricostruzione dell'ambizione dell'umano di cui ci stiamo occupando. Il Rinascimento europeo aveva a piccoli passi rimesso al centro l'opera di conoscenza della natura, con figure di spicco come Leonardo da Vinci, Giordano Bruno e Niccolò Cusano (Rossi 1962; Zellini 1980), e aveva posto il corpo umano come oggetto di indagine alla stregua di tutti gli altri. Emblematica in tal senso è l'accettazione diffusa del metodo anatomico e le progressive scoperte derivanti dal suo ampio impiego, in primis la scoperta della circolazione del sangue da parte di William Harvey, negli anni '20 del XVII secolo e la fondazione della 'iatromeccanica'<sup>17</sup>. Grazie a questa scoperta, il medico inglese propone l'immagine del corpo come una macchina idraulica che "batte" come un orologio (Wood 2002; Israel 2004: 64).

L'orologio assume in questo contesto un ruolo centrale nella figurazione del corpo umano a partire dal Rinascimento, già con le tavole anatomiche disegnate da Leonardo<sup>18</sup>, non solo per il funzionamento che permetteva una misurazione del tempo, già di per sé una svolta epocale, ma anche perché i primi automi meccanici furono le figure mobili sulle torri dell'orologio del Rinascimento. Il più impressionante di tutti per gli uomini e le donne del tempo era senza dubbio l'orologio di Strasburgo, costruito nel 1574, che presentava un gallo in ghisa come simbolo cristiano del rinnegamento di Gesù da parte di Pietro. Il gallo usciva a mezzogiorno accompagnato da altre dodici statue che raffiguravano gli apostoli, apriva il becco, tirava fuori la lingua, batteva le ali, sollevava le penne, tirava su la testa e cantava tre volte grazie all'impiego della nuova tecnica dei mantici e di una

---

<sup>17</sup> Anche detta "iatrofisica", la 'iatromeccanica' si basa sull'assumere come principi primi del funzionamento del mondo e del corpo umano quelli della meccanica della leva o della pompa idraulica. (Israel 2004: 68).

<sup>18</sup> È interessante notare come, se sul versante tecnico-scientifico Leonardo può ancora apparire come un uomo pre-rinascimentale, in quanto la sua ricerca, «pur piena di balenanti intuizioni e di geniali vedute, non oltrepassò mai il piano degli esperimenti *curiosi* per giungere a quella sistematicità che è la caratteristica fondamentale della scienza e della tecnica moderne» (Rossi 1962: 47), sia l'arte il terreno più fertile per una riflessione scientifica leonardesca nel vero senso del termine: «nei disegni di macchine come in quelli anatomici, Leonardo dava davvero un contributo decisivo all'invenzione di un metodo preciso di raffigurazione e descrizione della realtà. Il generico empirismo si fa qui sperimentalismo [...]» (Ibid.)



canna. L'orologio rimarrà funzionante fino al 1789, ispirando personaggi come Hobbes, Cartesio e Boyle, che rimasero affascinati da quell'esempio di meccanica avanzata (Heppenheimer 1985: 30-31).

Come sottolineato da Koyré (1948: 106-107), la messa a punto dell'orologio, non solo quelli delle torri campanarie o comunali ma anche quelli da tasca, contribuisce significativamente, insieme al perfezionamento dell'ottica e alla produzione di occhiali e lenti per l'indagine specifica, al passaggio dal mondo del pressappoco all'universo della precisione: l'accuratezza con cui si poteva misurare il tempo entra nella mentalità dell'uomo seicentesco e consente di sviluppare una visione metodica della sua impresa di conoscenza. L'orologio, dunque, assume un ruolo centrale nello sviluppo di una concezione moderna della scienza e diventa il simbolo di tutto ciò che è dato sapere della natura e dell'universo:

Indubbiamente, [l'orologio] è portato alla ribalta dall'immagine dell'universo come un sistema meccanico in cui il moto dei pianeti avviene secondo ritmi e tempi assolutamente regolari e armoniosi, come il concatenarsi dei movimenti delle ruote dentate di un orologio. D'altra parte, non è forse l'orologio l'immagine dello scorrere regolare, immutabile e assoluto del tempo, secondo principi di natura meccanica? E non fornisce l'immagine della struttura matematica dell'universo? (Israel 2004: 64)

Questa immagine contraddistingue non solo il versante scientifico della cultura del tempo, ma anche, ovviamente, quello religioso, legati entrambi a doppio filo e indagabili solamente se compresi come uno il terreno di validazione dell'altro. È così che Dio diventa il creatore di questa macchina perfetta che è l'Universo, il "Supremo Orologiaio" che ha messo insieme gli ingranaggi come un sapiente artigiano della creazione, da cui i moti e i funzionamenti della macchina partono e a cui fanno ritorno.

La figura centrale che raccoglie e sviluppa questa immagine, costruendo una visione dell'universo e delle sue creature come macchine o meccanismi e dando vita a quella corrente filosofica che prende il nome di meccanicismo razionalista, è quella di Cartesio. La chiave interpretativa della realtà e della scienza diventa quindi quella di leggere e interpretare i fenomeni naturali tramite i principi della meccanica. La macchina, intesa come assemblaggio di meccanismi e ingranaggi che funziona grazie ai principi del moto locale<sup>19</sup>, diventa non solo una via di fuga dalle spiegazioni magiche e dalle pratiche oscure adottate e studiate fin dall'antichità per comprendere la peculiarità del vivente<sup>20</sup>, ma

---

<sup>19</sup> Cartesio è passato alla storia, oltre che per aver reso possibile la commensurabilità tra matematica e geometria con l'invenzione del piano che porta il suo nome, anche per la formulazione della teoria fisica dell'urto, che presupponeva che tutte le forme di movimento si realizzassero grazie a un urto della materia. La visione entrerà in conflitto poi con la nozione di forza, e soprattutto di forza di gravità, elaborata da Newton, da cui verrà surclassata (Zellini 1980).

<sup>20</sup> Nello specifico, in più occasioni Cartesio si schiera apertamente contro il perseguimento della conoscenza attraverso "vie fuori dall'ordinario" a cui si erano affidati prima di lui maghi, naturalisti e alchimisti (cfr. *Descartes à Huygens*, in *Œuvres*, Paris 1949, p. 786). Il discorso viene coniugato anche all'interno di un'ottica intersoggettiva della scienza, non intesa in termini di raggiungimento congiunto di verità o conoscenza finale, ma come la capacità di assommare le esperienze per portare più prove alle intuizioni del maestro. Cfr. Rossi (1962: 115-119).

si erge a strumento di pensiero utile a comprendere la struttura dell'universo tutto e le singole unità che lo compongono. Nella formulazione cartesiana, l'uomo, dunque, oggetto tra gli oggetti, viene inquadrato nella sua componente meccanicista al pari di ogni altro ingranaggio nel grande orologio del creato.

Suppongo che il corpo non sia altra cosa che una statua o una macchina di terra, che Dio forma di proposito, per renderla la più simile a noi che sia possibile: di modo che, non soltanto gli dà all'esterno il colore e la figura di tutte le nostre membra, ma anche gli mette all'interno tutti i pezzi che sono necessari affinché funzioni, che mangi, che respiri, e infine che imiti tutte quelle nostre funzioni che possono essere supposte procedere dalla materia, e dipendere soltanto dalla disposizione degli organi. Noi vediamo orologi e fontane artificiali, mulini e macchine simili, che pur essendo state fatte soltanto da uomini, nondimeno hanno la forza di muoversi da sole in più modi diversi; e mi sembra che non saprei immaginare tanti tipi di movimenti, in questa che suppongo essere stata fatta dalle mani di Dio, né attribuirle tanti artifici, che si abbia motivo di pensare che ve ne possono essere ancora altri. (Descartes, *Œuvres et lettres*, Gallimard, Paris, 1953, p. 807).

In questo passo emerge chiara l'ispirazione di Cartesio alle macchine che aveva visto o di cui aveva sentito parlare nel corso della sua vita: statue semoventi, fontane prodigiose, orologi meccanici vengono presi a metafora del funzionamento del corpo umano, tanto da ritrovare la descrizione dei nervi come tubi cavi, che sembra richiamare il funzionamento della campanella per chiamare i domestici<sup>21</sup> (Cohen 1966: 64). L'unica differenza che va necessariamente sottolineata è che, diversamente da altre creature viventi come gli animali, gli uomini posseggono anche un'anima razionale, qualcosa che li contraddistingue nell'opera del creato e che li eleva in una posizione più vicina a quella divina. Viene formulato così il dualismo cartesiano, la netta contrapposizione tra corpo materiale e anima spirituale, *res extensa* e *res cogitans*, corpo e mente.

A lato della sua formulazione filosofica della teoria dell'universo come macchina, alcune fonti testimoniano anche del rapporto stretto di Cartesio con la pratica, che avrebbe portato alla realizzazione di alcune macchine antropomorfe. Si narra che, a seguito della morte prematura della figlia Francine, il filosofo avesse costruito una piccola bambola semovente, ma che questa fosse stata gettata poi in acqua dai marinai spaventati che lo scortavano nel suo viaggio verso la Svezia, l'ultimo della sua vita (Wood 2002). In realtà, non c'è alcuna prova della realizzazione materiale della bambola, ma numerosi scritti riportano i processi di ideazione della stessa insieme ad altre complesse creazioni meccaniche come un danzatore, un piccione capace di volare e un cane che insegue un fagiano (*Trattato sull'uomo*).

Le idee di Cartesio ispireranno una nutrita serie di riflessioni filosofiche e stabiliranno per secoli il ruolo della macchina come principale modello del mondo, la cui indagine svelerebbe allo scienziato

---

<sup>21</sup> «Questi nervi, secondo Cartesio, avevano il compito di azionare le valvole della testa, che a loro volta dovevano dirigere il flusso si "spirito animale" dal cervello ai muscoli, la ghiandola pineale, nella quale Cartesio faceva risiedere la mente, che aveva funzione di valvola principale, e poteva controllare tutte le altre». (Cohen 1966: 64)

i segreti del funzionamento dei suoi ingranaggi. Di specifico interesse per il nostro obiettivo di ricostruzione del concetto di macchina è il fatto che la filosofia cartesiana ispirerà direttamente anche la creazione di numerosi congegni che potessero riprodurre il comportamento e le fattezze umane o animali, realizzati sulla base del principio della visione delle creature come automi.

In particolare, il Settecento viene battezzato come il secolo degli automi, complice la diffusione e la rielaborazione delle idee cartesiane in chiave monista o materialista (Israel 2004: 66). L'attenzione per le macchine viene resa palese anche nella raccolta pubblicata nel 1735 in sei volumi da Jean-Gaffin Gallon, che descrive tutte le *Machines et inventions approuvées par l'Académie Royale des Sciences depuis son établissement*. In questi volumi si possono trovare le prime descrizioni delle macchine che al tempo si consideravano frutto di ingegno avanguardistico, soprattutto nel campo dell'automazione del calcolo, come l'*abaque rhabdologique* di Perrault, le *machines arithmétiques* di Lepine e di Pascal, e i diversi automi di Maillard, ma anche metodi per la costruzione di carrozze, sistemi di chiuse, mulini, macchine per produrre carta, arti artificiali e strumenti per fare la punta alle penne d'oca (Marchis 2014).

Esempio più celebre della sistematizzazione settecentesca del sapere è la celebre *Encyclopédie* redatta da Diderot e D'Alembert. È interessante notare come in questa sede l'automa sia definito come «engin qui se meut de lui-même, ou machine qui porte en elle le principe de son mouvement» (I, 896), non discostandosi dalla definizione originaria del termine ma concentrando l'attenzione sulla autonomia del moto prodotto dalla macchina. In questa rassegna vengono elencati anche i più famosi automi dell'epoca: rapisce l'attenzione dei due enciclopedisti l'opera di un famoso inventore, tale Jacques de Vaucanson<sup>22</sup>. La sua abilità di costruttore di automi passerà alla storia come strabiliante, e numerose sono le testimonianze della meraviglia di fronte alle sue costruzioni. Tra i tanti realizzati, due sono stati i più fortunati, forse proprio perché inseriti all'interno dell'*Encyclopédie*.

Il primo è un'anatra, forse l'automa più celebre costruito da Vaucanson, che riproduce le azioni di mangiare, bere e digerire. L'anatra era costruita seguendo un modello anatomico dettagliatissimo, che riproduceva la maggior parte delle ossa delle ali, consentiva diversi movimenti e garantiva l'esecuzione del compito una volta che fosse stata messa in moto. Inoltre, l'anatra aveva i meccanismi volontariamente lasciati scoperti a mostrare la complicatezza del suo funzionamento per accrescere le aspettative dello spettatore riguardo i suoi meccanismi interni. L'anatra, infatti, veniva fatta esibire da Vaucanson che la nutriva con del grano, per poi far vedere come il cibo fosse digerito dall'automa

---

<sup>22</sup> Tra le clamorose invenzioni di Vaucanson va annoverato anche un sistema di tessitura automatico, che ai suoi tempi non fece molta fortuna. Fu sfruttato poi da Joseph-Marie Jacquard il quale, all'inizio del XIX secolo, costruì un congegno che univa l'invenzione di Vaucanson a quella degli aghi e dei cartoni perforati di Basile Bouchon e Jean Baptiste Falcon, mettendo quindi in circolazione la prima macchina tessile programmabile. Sarà il modello da cui Charles Babbage trarrà ispirazione per la costruzione della macchina analitica, considerata un antenato prossimo del primo calcolatore (Losano 1990).

mostrando i suoi escrementi. In realtà, come si sarebbe scoperto in seguito, l'anatra non possedeva davvero la capacità di digerire il cibo, ma il materiale era già presente al suo interno e veniva mostrato all'occorrenza. Il secondo celebre automa costruito da Vaucanson annoverato nell'*Enciclopedia* è il "joüeur de tambourin", il suonatore di tamburello, facente parte di una serie di danzatori e suonatori artificiali costruiti da Vaucanson su ispirazione degli scritti di Cartesio. Questo specifico automa viene descritto capace di suonare una ventina di brani, minuetti, rigodoni o contradanze, e risulta di particolare interesse perché, sebbene possa sembrare facile riprodurre il meccanismo di suonare un tamburo rispetto ad altri strumenti, gli enciclopedisti si affrettano a spiegare che risulta uno dei più difficili: «Ce mouvement n'est pas un des plus aisés de la machine. Il est question de frapper tantôt plus fort, tantôt plus vite, & de donner toujours un coup sec, pour tirer du son du tambour. Cette mécanique consiste dans une combinaison infinie de leviers & de ressorts différens, tous mûs avec assez de justesse pour suivre l'air; ce qui seroit trop long à détailler» (Diderot e D'Alembert 1751, I: 897).

Un'ulteriore caratteristica dell'*Enciclopedia* settecentesca attira la nostra attenzione, in quanto viene inserito nel novero degli oggetti interessanti anche l'androide, definito come «automate ayant figure humaine & qui, par le moyen de certains ressorts, &c. bien disposés, agit & fait d'autres fonctions extérieurement semblables à celles de l'homme» (Ibid.: 448). È in questa voce che gli autori inseriscono una dettagliatissima descrizione dei meccanismi che animano un altro automa costruito da Vaucanson, "le flûteur automate". Il suonatore di flauto era un automa che riproduceva in modo accuratissimo le azioni di un abile musicista attraverso un sistema di tubi e cannule che incanalavano l'aria e, grazie alla riproduzione fedele delle dita di un suonatore umano che potevano così chiudere i fori del flauto in modo simile, questa veniva sfruttata per produrre dolci note di diverse melodie. La voce enciclopedica si dilunga in dettagliatissime spiegazioni di come sia stato possibile per il costruttore riuscire a replicare la sonatura del fluato finanche alle note più impegnative<sup>23</sup>, mostrando l'ammirazione degli intellettuali del tempo per gli automi.

In sostanza, l'*Enciclopedia* assume un ruolo centrale nella spinta che viene data alla visione meccanicistica dell'universo e dell'umano, e nel panorama filosofico si assiste a una continua rielaborazione delle tematiche inaugurate da Cartesio. In particolare, sotto l'influenza di un approccio dichiaratamente materialista, la fisiologia Settecentesca espelle l'anima spirituale elaborata da Cartesio con la nozione di *res cogitans* e la sostituisce con il cervello, reinterpretando di fatto le elucubrazioni cartesiane in chiave monistica. Figure chiave in questo processo sono Hobbes e Locke: il primo espone una visione in cui l'universo non è altro che il risultato del movimento dei corpi e in particolare lo è il pensiero, che viene visto come conseguenza di moti di parti del cervello e che porta

---

<sup>23</sup> Per una disamina approfondita delle creature di Vaucanson, si veda Losano 1990: 81-95.

a considerare la psicologia come una branca della meccanica; il secondo tenta di ridurre il pensiero a un insieme di sensazioni elementari, configurandolo come un aggregato di fatti empirici (Israel 2004). In questa cornice si inserisce anche il tentativo di Pascal di automatizzare il calcolo con la costruzione della sua celebre pascalina, poi perfezionata da Leibniz per arrivare a computare tutte e quattro le operazioni fondamentali della matematica (Davis 2000; Zellini 1980).

Altra figura centrale nell'operazione di mettere al centro lo studio dei meccanismi dell'uomo è quella del filosofo Julien la Mettrie, autore di un'opera provocatrice come *L'Homme-Machine* (1747) che estremizza le idee di Cartesio sul meccanicismo arrivando ad eliminare dall'universo tutti i fattori non meccanici e a trasformare il dualismo cartesiano in un monismo meccanico operante tanto nell'uomo quanto nelle altre creature. Come sottolineato da Cohen (1966: 65), per la Mettrie «non soltanto tutti gli animali sono organizzati come macchine, [...] ma lo stesso organismo umano, nella sua interezza, è soltanto una forma più perfetta dell'animale-macchina». Ritorna inoltre con forza l'immagine dell'orologio per spiegare il meccanicismo del corpo umano come assemblaggio di molle e congegni che si attivano reciprocamente senza la possibilità di stabilire un punto di inizio del processo osservato, avendo in sé stesso il principio di produrre il suo stesso movimento. Questa visione, a volte anche ostracizzata per la sua estrema, viene comunque accolta dalla cultura del tempo e contribuisce a stimolare l'immaginazione dei costruttori di automi.

Uno tra i seguaci della filosofia di la Mettrie, importante per il ruolo che le sue opere assumeranno nel secolo successivo, è l'orologiaio Pierre Jacquet-Droz. Insieme a suo figlio, costruirà tre automi che diventeranno molto popolari nell'Europa della metà del XVIII secolo: la suonatrice, il disegnatore e lo scrivano. I meccanismi di natura pneumatica o idraulica di questi automi riuscivano a dare l'impressione, una volta avviati, che le macchine svolgessero autonomamente quelle attività che avevano fatto loro guadagnare il proprio nome: la suonatrice riusciva a riprodurre cinque melodie diverse all'organo, ma producendo il suono di un flauto; il disegnatore è di dimensioni leggermente maggiori ed esegue cinque disegni, fermandosi di tanto in tanto a soppesarne i dettagli e a liberare il foglio dalle impurità lasciate dalla matita; lo scrivano è una figura di bambino dalle fattezze delicate che traccia con una penna d'oca brevi frasi su un foglio (Losano 1990: 99-100). Gli automi erano parte di una campagna pubblicitaria per la vendita di orologi ai nobili dell'Europa del XVIII secolo e vennero esposti a Neuchâtel, in Svizzera, dove si narra Mary Shelley li avesse visti durante un soggiorno in montagna con alcuni suoi amici scrittori, lo stesso soggiorno in cui venne abbozzata la storia di un dottore che aveva lo scopo di creare la vita dalla morte, poi pubblicata nel 1818 con il titolo *Frankenstein, o il moderno Prometeo*.

Nello stesso periodo degli automi di Jacquet-Droz, un altro automa sorprende le corti europee, e sarebbe passato alla storia come una delle macchine più famose di tutti i tempi. È il giocatore di

scacchi costruito dal barone von Kempelen nel 1769, assemblato per stupire l'imperatrice Maria Teresa d'Austria e finito a popolare le pagine di matematici, con lo scopo di svelarne l'arcano, e di autori di pregio come Edgar Alla Poe. La storia della sua ideazione è alquanto particolare, ma vi si rintraccia la stessa curiosità per le leggi fisiche e l'applicazione di una concezione meccanicista tipica di tutti gli automi del Settecento. Invitato ad assistere alla dimostrazione di esperimenti sulla forza magnetica, il barone rimase affascinato tanto quanto l'imperatrice, a cui promise di realizzare qualcosa di ancor più sorprendente. In poco più di sei mesi, von Kempelen presenta alla sovrana un automa che passerà alla storia come il "turco meccanico", che aveva una straordinaria abilità nel gioco degli scacchi. L'abbigliamento esotico e la postura rilassata hanno contribuito a rinforzare l'interesse per questa strana figura, che girerà il mondo grazie al suo padrone e costruttore battendo sfilze di audaci sfidanti. Data la straordinaria abilità dell'automa nel giocare a scacchi, diversi commentatori del tempo ipotizzarono che necessariamente dovesse esserci una persona nascosta all'interno del turco, permettendogli di raggiungere lo scopo per cui era stato costruito. Tuttavia, è stato per molto tempo difficile capire sia come, sia dove potesse nascondersi tale persona. Le abilità del turco meccanico, poi, non si esaurivano con gli scacchi, aumentando ancor più l'alone di mistero: a seguito della partita, durante la quale il barone dava la carica con una chiavetta all'automa ogni dieci o dodici mosse, la scacchiera veniva sostituita con una tavola su cui erano incisi numeri e lettere, e all'automa poteva esser rivolta qualsiasi domanda a cui questi trovava risposta indicando le incisioni sulla tavola, alimentando ipotesi che richiamavano la natura magica o mistica dell'automa. Dopo la morte del barone von Kempelen, l'automa fu venduto all'inventore viennese Johann Nepomuk Maelzel, che continuò le tournée europee fino a portare il turco a giocare anche con Napoleone, spostandosi poi in America e ai Caraibi. L'automa godette per un periodo della fama degli inizi, mai più raggiunta dopo che numerosi matematici e inventori avevano svelato il necessario trucco dell'uomo minuto nascosto al suo interno, e fu distrutto da un incendio nel museo di Philadelphia in cui era esposto nel 1854 (Losano 1990: 100-116).

Come nel caso di alcune delle spiegazioni elaborate per il funzionamento del turco meccanico, la componente magica non ha mai abbandonato del tutto il campo della realizzazione dei doppi, anche quando l'ambizione di realizzarne di sempre più simili all'umano era stata affiancata dalla scommessa epistemologica di costruirne per comprendere l'essenza dell'universo. Nel periodo tra il XVII e il XVIII secolo, l'ambizione dell'umano che stiamo indagando si mescola nuovamente alla magia, assumendo nuovi tratti e incarnandosi nel mito del Golem. La versione più conosciuta oggi è quella offerta dal racconto di Jakob Grimm, scritto nel 1808, e dalle sue varie riprese a seguito della pubblicazione. Tuttavia, il mito del Golem, la statua di pietra vivente, affonda le sue radici nella tradizione ebraica: 'golem', infatti, in lingua ebraica sta per 'massa informe' e viene spesso accostata

alla sostanza da cui Dio crea Adamo, con riferimento a qualcosa di incompleto e ancora da modellare. La leggenda da cui uno dei fratelli Grimm recupera il mito del Golem comincia a diffondersi intorno alla metà del Seicento e conserva la stessa radice ebraica e talmudica. Il racconto narra del capo rabbino di Praga, Giuda ben Loew, che si voleva anche amico di Keplero e Tycho Brahe, il quale per proteggere la sua gente dai pogrom si recò con i suoi assistenti sulle rive della Moldavia e lì raccolse dell'argilla con cui plasmarono una figura umana<sup>24</sup>. Mentre un assistente girava intorno alla statua sette volte da destra verso sinistra, l'altro faceva lo stesso in senso contrario e il rabbino pronunciava formule magiche: questo fece spuntare capelli e unghie al Golem e, una volta che anche il rabbino Loew aveva impresso sulla fronte della statua il Sacro Nome, questo prese vita. Al gigante mancava la parola, ma possedeva una forza sovraumana che lo aiutò a difendere gli ebrei dalle persecuzioni seguendo gli ordini del suo creatore. Come accade in molte delle narrazioni riguardo alle creature artificiali, però, anche il Golem sfugge al controllo del suo artefice e gli si ribella scatenando la sua furia distruttiva. Lo stratagemma adottato dal rabbino è quello di cancellare con un inganno la parola ebraica che il Golem portava scritta sulla fronte, rendendolo di nuovo un cumulo di argilla (Heppenheimer 1985: 27-29). Le varianti della fine del Golem sono molte, ma tutte mettono in risalto la mente lucida del costruttore e la sua responsabilità nel dover fermare la furia che la creatura da lui portata in vita aveva liberato. Per alcuni, inoltre, la morale della leggenda non è tanto quella di non spingersi a replicare solo ciò che Dio era riuscito a fare, ma di non peccare di *idolatria* portando sullo stesso piano le proprie creature artificiali imperfette e l'opera di Dio, che invece è completa nella sua finalità (Israel 2004: 36-37).

Come testimonia la ripresa moderna<sup>25</sup>, il mito del Golem diventa un *leit motif* della narrazione riguardo al rapporto tra creatura e inventore e la sua rielaborazione agli inizi dell'Ottocento costituisce un diretto antenato di molti dei racconti di fantascienza di cui abbiamo parlato all'inizio di questo capitolo<sup>26</sup>. Ora, quello che si vuole approfondire è una riflessione sul ruolo che le macchine automatiche elaborate nel corso dei tre secoli che vanno dall'inizio del Cinquecento alla fine del Settecento hanno avuto per lo sviluppo e l'evoluzione dell'ambizione dell'umano.

Se abbiamo definito le macchine che venivano prodotte nel periodo che va dall'antichità al Medioevo come *macchine idraulico-pneumatiche*, perché identificate dal loro funzionamento basato su pesi o sull'utilizzo di forze naturali come l'acqua per infondere l'apparenza del movimento

---

<sup>24</sup> Si noti anche in questo caso il riferimento al mito di Prometeo (*infra* § 1.1.2).

<sup>25</sup> Tra tutti i richiami moderni alla leggenda del Golem, uno dei più interessanti è senza dubbio il volume di Norbert Wiener *Dio & Golem S.p.a. Cibernetica e religione* (1966), insieme ad alcune riprese ancora più recenti in robotica (Pollack e Lipson 2000).

<sup>26</sup> Anche la semplice enumerazione dei racconti che includono macchine autonome o androidi nel periodo che abbiamo preso in esame andrebbe oltre lo scopo della nostra argomentazione. Per una disamina approfondita si veda Losano 1990; Wood 2002.

autonomo, le macchine di questo periodo possono essere definite un'*orologeria idraulico-meccanica* (Israel 2004: 65): il meccanismo dell'orologio ben si addice a riassumere la visione della macchina autonoma di questo periodo storico-filosofico, che non si aiuta solamente con forze di moto naturali, e le concezioni idrauliche del corpo umano come quella proposta da Harvey fanno propendere per una descrizione dell'ambizione dell'umano sotto questa metafora. È nel connubio tra meccanicismo cartesiano di stampo razionalistico e le scoperte relative al funzionamento del corpo umano, come quella della circolazione del sangue, che si può coniugare la visione l'orologio come insieme di ingranaggi azionati dall'urto, attraversato a sua volta da flussi idraulici e messi in moto dall'interno. Ora, come questo ci permette di avanzare nell'indagine dell'ambizione dell'umano?

A questo scopo, è conveniente rivolgere l'attenzione al ruolo che Cartesio riserva alle macchine vere e proprie, quelle realizzate al suo tempo, nella costruzione del suo "universo-macchina". Infatti, per Cartesio le macchine non sono altro che uno strumento euristico, un tramite attraverso cui rinforzare l'uso personale dello scienziato delle «idee chiare e distinte presenti nel nostro intelletto» (Cartesio 1644). È in questa sede che la macchina diventa quindi metafora tanto dell'universo quanto dell'uomo che lo abita, ma esclusivamente funzionale alla comprensione più alta dell'ordine globale, legato alla metafisica più che all'empirismo. Le tesi radicali di Cartesio in merito alla natura meccanicista del mondo, che lo portano a rintracciare una completa identità tra artificiale e naturale (per cui tutto ciò che è artificiale è anche naturale), conducono a una profonda riconfigurazione del concetto generale di natura, che «non appare più contesta di forme e di essenze cui ineriscono le "qualità", ma di fenomeni quantitativamente misurabili» (Rossi 1962: 144). Di fatto, il processo di conoscenza della realtà si esplicita nella identificazione delle qualità in termini matematizzabili e quantitativamente misurabili: tutto ciò che sfugge a tale processo viene espulso dalla conoscenza formalizzata e formalizzabile. In questo caso, il concetto di macchina diventa funzionale a identificare tutto ciò che può essere raccolto sotto l'etichetta del misurabile, del quantificabile e quindi, in sostanza, del riproducibile (Israel 2004, 2015; Ceruti 1985). Ma il percorso non è ancora a doppio senso: il pensiero cartesiano non vede ancora l'uomo-macchina come un'unità inscindibile i cui termini sono perfettamente reversibili; l'uomo non è ancora completamente inquadrabile sotto lo sguardo di questa metafora, come testimonia il dualismo che da Cartesio prende il nome e che pone uno iato tra materia e anima. Alcune interpretazioni radicalmente materialiste sorgeranno poi con le letture tardo settecentesche della filosofia di Cartesio, come quella di la Mettrie, ma anche qui il paragone tra l'uomo e la macchina non assumerà i tratti di un'impresa di conoscenza biunivoca: la macchina, sì, esemplifica il funzionamento del corpo umano, perché tutto è macchina compreso il pensiero, ma non si crede che costruendo macchine si possa arrivare a una comprensione più profonda di come sia fatto l'umano. Sebbene, quindi, il legame tra i due poli della metafora si saldi con



l'affermarsi del meccanicismo, garantendo i termini di commensurabilità tra l'uomo e la macchina, l'orologio idraulico-meccanico che costituisce il modello della macchina per quel tempo non produce ancora la possibilità di nuova conoscenza, quanto solamente materia di osservazione.

È interessante notare come sia a partire dalla formulazione del dualismo cartesiano che si produce una biforcazione dei percorsi intrapresi dall'ambizione dell'umano di replicarsi. Infatti, mentre fino a Cartesio il sentiero si era mostrato pressoché uniforme, prendendo correnti e abbracciando arti diverse ma costituendosi come un'operazione unitaria nella storia umana – quella di replicare le sue abilità in qualche maniera – a seguito della spaccatura tra corpo e mente da lui individuata l'ambizione dell'umano si scinde. Da una parte, è rintracciabile la continuazione dell'impresa precedente, in ambiti di applicazione come l'arte, la letteratura, alcune forme di ingegneria e di intrattenimento, che insistono nel realizzare artefatti completi che riproducano l'umano in qualche sua forma. Dall'altra, si assiste alla proliferazione di un'altra linea di indagine, basata sulla realizzazione di artefatti che mimino le strutture logiche del linguaggio e della matematica, ma non le fattezze umane. Alcuni esempi sono già stati citati, come la macchina per pensare di Raimondo Lullo o la Pascalina di Blaise Pascal, e avranno numerosi eredi nei secoli a venire, come ad esempio la macchina calcolatrice di Leibniz o quella di Charles Babbage. A partire da Cartesio, ovvero dal momento in cui la mente si stacca dal corpo e il pensiero si autonomizza dalla sostanza materiale, è possibile pensare alle strutture del pensiero come autosufficienti, indagabili e riproducibili per mezzo della macchina. Si apre così un nuovo scenario, che potremo definire proto-funzionalista (Putnam 1960, cfr. §1.3.1), che darà i suoi frutti più maturi solamente nel XX secolo con l'invenzione dell'Intelligenza Artificiale e la formalizzazione della logica computazionale (Davis 2000).

Dalla metà del Settecento, accanto al meccanicismo, si va poi evolvendo un nuovo sguardo alla natura che riprende un filone vitalistico-organicistico che finirà per caratterizzare il tardo Settecento e dare vita all'indagine sulla biologia nel secolo successivo. L'esigenza di creare un campo di studi indipendente dalla fisica per studiare il fenomeno della vita nasce dal riconoscimento dell'impossibilità di spiegare esaurientemente i fenomeni organici all'interno di uno sguardo causale meccanicistico (Bernardi 2014). Come sottolineato egregiamente da Israel, il parziale fallimento del progetto della fisica cartesiana è legato

[...] alla inadeguatezza a descrivere i fenomeni fisici, alla sua inconsistenza (in particolare per quel che riguarda la teoria dei vortici), alla sua incapacità di realizzare concretamente il programma di *matematizzazione della realtà*, e quindi tradurre dalle intenzioni ai fatti il progetto di descrivere matematicamente i fenomeni fisici, in modo da poterli prevedere. (Israel 2015: 4)

Il problema della irriducibilità dei fenomeni fisici al programma del meccanicismo causalistico cartesiano fondato sulla teoria degli urti, d'altronde, aveva trovava già il suo antagonista in un'altra

linea di ricerca, che va da Galileo a Newton. Rintracciare le differenze tra queste due linee di ricerca, o i possibili margini di congruenza, va di molto oltre allo scopo del presente paragrafo<sup>27</sup>, ma su un punto è opportuno soffermarsi. Tale punto riguarda il perché, nonostante il cartesianesimo abbia avuto vita breve nel campo nella fisica, soppiantato alacrementemente dalle teorie newtoniane delle forze basata sull'empirismo e non sul razionalismo, questo abbia ancora una profonda influenza sul modo in cui concepiamo oggi le possibilità di realizzazione dell'ambizione dell'umano.

Mentre Cartesio, con il suo approccio totalizzante volto congiungere metafisica e fisica, parla esplicitamente delle macchine come strumento adeguato a comprendere il funzionamento del mondo, Newton si "limita" a formulare leggi riguardo all'ordine dell'universo, senza saldare le sue riflessioni al rapporto tra anima e corpo o coniugando metafore attraverso cui era possibile spiegare l'emergere del libero arbitrio o le possibili connessioni in termini di pensiero individuale. In questo senso, Newton può fornire un rilancio della teoria vitalista e organicista che non confina i fenomeni organici all'interno della spiegazione causalistica meccanicista, ma non fornisce appiglio per lo sviluppo di un pensiero inerente al modo in cui la struttura dell'umano possa essere replicata in qualche modo. Si può dire che questa verrà sviluppata da Leibniz, che resterà scettico nei confronti di una completa riproducibilità dell'intelletto umano nelle macchine e che propenderà per una visione più vitalista, anche se contribuirà in modo sostanziale a collocare la metodologia meccanicista al cuore del metodo scientifico (Israel 2015).

Nella fase in cui la rivalutazione di un'impresa di conoscenza, spinta da una parte dalla possibilità di matematizzazione dell'universo, dall'altra sfruttando le nuove idee sulle leggi fisiche, la rilettura del mito di Prometeo ci dimostra il cambiamento culturale in opera in quel periodo. Se finanche i primi pensatori cristiani non cessano di condannare la *hybris* della 'mala curiositas' attribuita a Prometeo, nel Seicento questo assume una valorizzazione positiva: da titano, creatura vicina agli dei, Prometeo si fa uomo e diventa simbolo di una nuova riconcentualizzazione del perseguimento della conoscenza, il suo osare e travalicare i limiti imposti dalla divinità diventano un pregio dello scienziato «per non aver [...] lasciato nulla di intentato nella ricerca della verità» (Battistini 2015: 198). Preso come eroe della conoscenza da figure come Erasmo da Rotterdam, Bacone e Galileo, Prometeo arriva qui a esemplificare l'impresa della scienza: la sfida all'ignoto, all'*auctoritas*, l'emblema di un paradigma indiziario che identifica la conoscenza come la fine di un percorso investigativo fatto di tracce e tranelli, fino a incarnare l'ideale stesso di progresso (Rossi 1962, Battistini 2015).

---

<sup>27</sup> Per un'approfondita disamina della relazione tra le idee di Cartesio e Newton: Israel (2015: 1-90); per una riflessione tecnica sul ruolo della filosofia naturale di Newton nel panorama filosofico del Novecento: Koyré (1965).

Rimarrà sempre, però, quella doppia accezione: chi sfida i limiti imposti dalla natura corre il rischio, nella migliore delle ipotesi, di farsi rodere il fegato; nella peggiore, va incontro ai pericoli aperti dalla nuova scienza.

### 1.2.2 *Macchine elettriche*

Figlia del secolo delle macchine, nel 1818 viene pubblicata un'opera che fa da ponte concettuale e immaginifico tra le macchine del secolo precedente e quelle che verranno costruite da quel momento in poi, raccogliendo le concezioni del passato per rilanciarle in un futuro non più determinato solamente dalla meccanica, ma anche da nuove forme di energia come l'elettricità. La storia del dottor Frankenstein, infatti, e della sua creatura mostruosa, possono raccontarci dello *Zeitgeist* dell'inizio dell'Ottocento e del modo in cui le scoperte scientifiche legate alle nuove forze e alle possibilità da queste aperte avevano portato linfa alla creatività, permettendo di immaginare nuove realizzazioni atte ad alimentare l'ambizione dell'umano.

Brevemente, la celebre storia racconta della volontà del dottor Frankenstein di ricreare la vita a partire dalla morte: lo scopo dello scienziato era quello di mettere insieme pezzi di cadaveri, trafugati da cimiteri e obitori, per poi farli tornare in vita attraverso uno strumento che convogliasse l'elettricità per trasmetterla all'assemblaggio anatomico. La creatura, che effettivamente prende vita, si dimostra però da subito "non umana", ripugnante alla vista e incapace di moderare la sua spropositata forza. Il rapporto travagliato che nasce dal rifiuto del padre-dottore della creatura porterà all'innescare di una serie di vendette e ritorsioni che non si risolveranno in un lieto fine, conducendo entrambi i protagonisti a soccombere sotto il peso delle loro colpe.

Il racconto è iconico e interessante in questa sede per una serie di ragioni. La prima tra tutte è rintracciabile già nel sottotitolo dell'opera: il *moderno Prometeo*. In quest'opera, infatti, è possibile registrare la coesistenza delle due interpretazioni che abbiamo messo in luce nei paragrafi precedenti: Mary Shelley, autrice del romanzo, rende perfettamente conto della doppia natura del mito di Prometeo, tenendo insieme, da una parte, il coraggio e la capacità di indagare i fenomeni della natura, di dare seguito alla curiosità sulla base di una scienza; dall'altra, il pericolo che da questa impresa deriva, ciò che si prospetta quando si affronta l'ancora sconosciuto e che condanna a subire le conseguenze delle proprie azioni.

In seconda sede, l'opera assume valore poiché mette in luce l'interesse espresso nei confronti delle nuove scoperte scientifiche, che si dimostra sfruttabile anche in domini non immediatamente attigui a quelli in cui le scoperte vengono sviluppate, come la letteratura. In particolare, il dottor Frankenstein, oltre a vaste conoscenze in biologia, chimica e anatomia, dimostra uno spiccato interesse per l'elettromagnetica, e soprattutto per il galvanismo. In alcuni esperimenti datati 1780, lo

scienziato bolognese Luigi Galvani aveva dimostrato che, facendo passare un flusso di corrente elettrica attraverso la zampa di una rana (staccata dal corpo), si causava una contrazione del muscolo (Boden 2006: 107). L'idea era che, visto che era già di pubblico dominio la relazione tra i nervi e la contrazione muscolare, l'impulso nervoso stesso fosse elettrico. Questo influenzò direttamente l'autrice del romanzo, la quale era bene a conoscenza degli esperimenti di Galvani prima e di Aldini poi, ispirando quindi la creazione del dottor Frankenstein (Piccolino e Bresaola 2003).

In generale, l'avanzamento tecnologico di cui si fa testimone il XIX secolo, oltre che a portare un incremento nella produzione industriale e a quella rivoluzione che ne avrebbe preso il nome, avrebbe dato i suoi frutti anche nel campo della realizzazione di doppi e copie con lo scopo di riprodurre l'essere umano. Ovviamente, la spinta alla costruzione di automi meccanici, soprattutto che riproducevano suonatori, danzatori o fanciulle graziose, non si esaurì nel Settecento, ma anzi crebbe notevolmente nell'Ottocento di pari passo all'interesse mostrato anche su vasta scala dal pubblico e in corrispondenza dell'affinamento delle tecniche di realizzazione e all'industrializzazione che ne permetteva l'ampia diffusione. Quello che è interessante indagare in questo frangente, tuttavia, riguarda un nuovo tipo di macchine, che sarebbe stato di ispirazione per la generazione successiva di psicologi e ingegneri, i quali avrebbero sviluppato un utilizzo metaforico ancora più dettagliato della macchina umanoide.

Come già accennato, il ruolo dell'elettricità e dei concomitanti sviluppi della teoria dell'elettromagnetismo contribuiscono in gran misura a sviluppare nuove idee per i concetti di macchina vivente. In particolare, il parallelismo tra la corrente elettrica e il funzionamento del corpo umano, già visto in Galvani, si pone al centro di una nuova formulazione di macchina vivente. Ad esempio, questa si manifesta in una riconfigurazione dello studio del cuore, che non può essere più visto solamente come una pompa idraulico-meccanica, ma viene inquadrato attraverso la sua attività elettrica. Nascono così gli elettrocardiogrammi come metodo di misurazione dell'attività cardiaca (Israel 2004). La letteratura, come abbiamo visto già testimone affascinata dalle potenzialità dell'elettricità, si popola di automi alimentati da questo tipo di energia. Nel corso di tutto il XIX secolo, ma anche per gran parte del XX, l'energia elettrica nei suoi più svariati impieghi diventa uno dei mezzi preferiti per donare la vita alle creature artificiali: dalla intrigante protagonista del romanzo *Eva Futura* di Villiers de l'Isle-Adam (1885) agli androidi di Asimov e Dick, tutte queste creature sono mosse da forze elettriche che consentono ai loro corpi di muoversi e comportarsi come esseri umani.

In secondo luogo, il riferimento va al nuovo tipo di macchine a vapore, che genera la metafora del corpo umano come produttrice di calore seguendo le teorie termiche e termodinamiche, le quali informano anche l'interesse medico per il metabolismo e il conteggio delle calorie. Lo sguardo,

dunque, si concentra sull'indagine e la comprensione degli scambi di carattere energetico-calorico che presiedono al funzionamento della macchina vivente (Israel 2004: 74). In contemporanea, anche nuove scoperte in campo chimico contribuiscono a informare la concezione del corpo umano e del suo conseguente meccanismo di funzionamento: i processi chimici sembrano essere orizzontali rispetto a tutte le forme di vita, sembrano regolarne ogni aspetto, dimostrandosi come il meccanismo determinante di ogni fenomeno dinamico. Anche in questo caso, la letteratura non fatica a raccogliere il seme della scoperta scientifica e sviluppa questa tematica in diversi racconti, di cui forse il più celebre è *Lo strano caso del dottor Jeckyll e Mr Hyde* (Stevenson 1886).

Come si sarà notato, gli esempi pratici della realizzazione dell'ambizione dell'umano in questo periodo storico sono più scarsi. Le motivazioni a cui è possibile ricondurre tale scarsità sono almeno due: una socioculturale ed economica, dovuta allo sviluppo della tecnologia industriale, e una filosofica. Sul piano socioculturale, la rivoluzione industriale ha convertito l'abilità dei costruttori di automi settecenteschi nella realizzazione di macchinari per la produzione: le macchine non hanno solamente il compito di divertire e sorprendere, ma devono essere utili, ovvero produrre beni economici, affermando quindi il dominio utilitaristico dell'umano sulla natura. Come sottolineato da Losano:

Nell'era della rivoluzione industriale, il meccanismo d'orologeria che muove gli automi è ormai una tecnologia matura, costantemente migliorata ma, ormai, non più innovativa. L'attenzione degli studiosi si concentra invece ossessivamente sul motore primario, destinato a rivoluzionare la società. Infatti, la meccanica tradizionale permette di costruire ormai qualsiasi macchina. Però solo il motore primario – la macchina a vapore, il motore a scoppio o elettrico – permette di muovere a centinaia e a migliaia queste macchine, moltiplicando per cento e per mille la forza fisica del medesimo uomo. (Losano 1990: 135).

Per questo, lo sviluppo della costruzione di doppi e riproduzioni dell'umano si biforca, rimanendo comunque confinata a domini economici: da una parte, viene portata avanti la costruzione degli automi destinati alle corti e ai nobili europei, con esempi di alta gioielleria e automazione negli automi gioiello di Fabergé e nelle bambole animate; dall'altra, la tecnica di riproduzione dell'umano trova sede nelle officine dei meccanici ottocenteschi, che non costruiscono più fedeli o approssimate rappresentazioni dell'umano a scopo di divertire o incantare il pubblico, ma osservano l'operaio e cercano di progettare e costruire strumenti utili a supportare, o anche sostituire, il suo lavoro<sup>28</sup>.

D'altra parte, sul piano filosofico il romanticismo portò a una battuta d'arresto per quanto riguarda la concezione della totale riproducibilità dell'umano in una macchina nella prima metà dell'Ottocento. I principi vitalisti tipici di questa corrente di pensiero, infatti, si ponevano in aperto

---

<sup>28</sup> La conversione della tecnica di riproduzione dell'umano a fini utilitaristici è sapientemente colta dal fisico von Helmholtz, che già nel 1854 scriveva: «Oggi infatti non cerchiamo più di costruire macchine che compiano mille diverse azioni di *un unico* uomo, ma al contrario esigiamo che la macchina compia *un'unica* azione, però sostituendo *mille* uomini». (von Helmholtz 1854: 140).

contrasto con una visione meccanicista della natura e degli organismi, preferendo altresì una visione organicista e incentrata sul rapporto con l'ambiente (Israel 2004, 2015). Tuttavia, gli sviluppi che in quel momento accostavano il funzionamento del corpo umano come meccanismo e le neonate indagini sul sistema nervoso degli ultimi anni dell'Ottocento avrebbero influenzato enormemente l'ideazione di macchine che imitassero la struttura cerebrale. In particolare, gli iniziatori di uno studio approfondito e sistemico del sistema cerebrale furono Camillo Golgi e Santiago Ramon-y-Cajal, che impiegarono nuove tecniche di contrasto e le innovative strumentazioni al microscopio per arrivare a vedere, definire e indagare quella che ancora oggi è considerata l'unità fondamentale del sistema cerebrale: il neurone. Fu grazie a questo tipo di indagini che fu possibile sostenere l'idea che il funzionamento del sistema nervoso centrale potesse essere considerato come una rete di comunicazione fra i neuroni (Changeux 1983).

L'Ottocento, dunque, è stato un secolo fondamentale per l'evoluzione dell'ambizione dell'umano, scarico se si guarda alle realizzazioni pratiche, ma ricco se ci si concentra sulla sua evoluzione concettuale e fantascientifica. Tuttavia, ancora qui è difficile rintracciare uno scarto con le idee riguardo al ruolo della macchina rispetto ai due secoli precedenti. Nell'Ottocento, ma anche nei primi anni del secolo successivo, la macchina serve ancora a realizzare le capacità tecniche dell'umano, a disvelare i segreti della natura (come, ad esempio, attraverso l'uso del microscopio), a potenziare le capacità umane (come nel caso delle macchine industriali capaci di sostituire gli operai), ma non ancora a capire qualcosa in più sulla natura della mente umana. Tuttavia, lo sviluppo della tecnica della fine del secolo ha avuto un ruolo fondamentale nell'emersione di una nuova prospettiva sulla scienza, che non viene più vista solamente come ricerca della verità: i nuovi problemi sorti all'interno del panorama scientifico, come quelli legati alla biologia, avevano reso necessaria una riconfigurazione del paradigma della conoscenza, facendo emergere questioni complesse che sarebbe stato possibile affrontare solamente attraverso il supporto di una tecnologia. Nell'interpretazione di Israel, «si può ben dire che questi sviluppi segnano il passaggio dalla tecnica alla *tecnologia*, intesa come *la tecnica basata sulla scienza*» (Israel 2015: 221).

### 1.2.3 Macchine cibernetiche

È proprio a partire dal Novecento che si vede un arricchimento di prospettive, una saldatura degli estremi: se la macchina era stata, almeno a partire da Cartesio, uno strumento attraverso cui capire il mondo, è soltanto inoltrandosi nel XX secolo – con un'idea di macchina vivente dalle caratteristiche principalmente fisico-chimiche, composta da strutture molecolari elementari, governata da principi adattivi e auto-adattivi ripresi dalla teoria darwiniana dell'evoluzione – che sarà possibile concepire il ribaltamento e il ritorno della metafora della macchina sull'umano. Con l'avvento della cibernetica

negli anni '40 del '900, poi, la macchina è diventata definitivamente il termine di paragone attraverso cui raggiungere una conoscenza profonda della natura e della struttura del comportamento e del corpo umani.

È importante, allora, rintracciare le caratteristiche della corrente cibernetica, che acquista una sua identità principalmente grazie all'opera di Norbert Wiener (1948) e di personaggi di spicco come Arturo Rosenblueth (Rosenblueth et al. 1943), William Ross Ashby (1946), e Warren McCulloch (McCulloch e Pitts 1943), e che influenzerà il modo in cui la mente viene concepita in quanto macchina fino ad oggi<sup>29</sup>.

La cibernetica raccoglie una serie di ricerche interdisciplinari sviluppate da ingegneri, psicologi, logici, fisici, matematici e teorici dei sistemi, aventi in comune l'obiettivo di avvicinare, fino a unificare, lo studio di sistemi naturali e sistemi artificiali, con la scommessa di poter replicare nella macchina il meccanismo di funzionamento di alcuni processi distintivi della vita. Il termine cibernetica viene coniato da Norbert Wiener nel 1947, attraverso un calco dal greco κυβερνήτης, *kybernētēs*, ovvero “pilota di navi”, ad indicare lo studio unitario dei processi riguardanti il controllo e la comunicazione nell'animale e nella macchina (Wiener 1948)<sup>30</sup>. Nonostante le differenze che distinguevano i ricercatori all'interno delle loro ricerche, il gruppo cibernetico si identificava per il loro interesse nello studio dei meccanismi di controllo dei sistemi viventi, con particolare attenzione allo studio del sistema nervoso e dei comportamenti degli organismi biologici. L'obiettivo condiviso era quello di riuscire a spiegare i meccanismi di pensiero e di comportamento degli oggetti osservati adottando un approccio fisicalista del pensiero. Si possono riassumere i principi fondanti della cibernetica i due assunti:

1. Pensare è una forma di computazione. Il calcolo richiesto non è l'operazione mentale di un essere umano che manipola simboli applicandovi regole come quelli della addizione o della moltiplicazione; invece si tratta di quello che una particolare classe di macchine fa – macchine tecnicamente concepite come “algoritmi”. In virtù di ciò, il pensare raggiunge l'ambito del meccanico.

2. Le leggi fisiche possono spiegare perché e come la natura – in alcune delle sue manifestazioni, non limitate esclusivamente al mondo umano – ci appare contenere significato, finalità, direzionalità, e intenzionalità.

---

<sup>29</sup> Si procederà a illustrare alcune delle condizioni per l'emersione della robotica in ambito cibernetico e pre-cibernetico, ma non sarà possibile esaurire in questa sede il discorso sulle influenze della cibernetica sulla nascita della robotica, né tanto meno sullo sviluppo delle scienze cognitive e dell'Intelligenza Artificiale. La ricostruzione storico-filosofica ed epistemologica del panorama pre-cibernetico, senza il quale l'emersione né della cibernetica stessa né della robotica sarebbero state possibili, è meravigliosamente illustrato in Cordeschi (1998). Le sfide lanciate dalla cibernetica come epistemologia sono discusse e sviluppate in Somenzi (1991). Sull'influenza della cibernetica per la creazione dell'impalcatura epistemologica delle scienze cognitive simboliche e subsimboliche: Dupuy (1994). Al testo capitale di Boden (2006) si fa riferimento per lo sviluppo della metafora della mente come macchina. Per i più recenti sviluppi dell'Intelligenza Artificiale e sulla loro natura ibrida simbolico-connessionista: Andler (2023).

<sup>30</sup> Il termine “cibernetica” viene ripreso da Wiener per ricordare uno dei componenti centrali del motore a vapore di James Watt, basato a sua volta sul meccanismo di controllo dei mulini a vento (Mataric 2007: 8).

[I fondatori del movimento cibernetico] ritenevano possibile costruire una teoria scientifica fisicalista della mente, e così risolvere l'antico problema filosofico del rapporto tra mente e materia. Si ritenevano in grado di riconciliare il mondo del significato con il mondo delle leggi fisiche. Grazie a loro, la mente avrebbe finalmente trovato il suo posto di diritto nella natura. (Dupuy 1994: 34)

Per attuare questo processo, uno sguardo privilegiato era riservato al rapporto che l'organismo/la macchina aveva con l'ambiente in cui era immerso. In sostanza, descrivere il rapporto stretto che si realizzava tra organismi biologici e il loro proprio ambiente era lo scopo primario della cibernetica, congiuntamente al proposito di formalizzare la descrizione e di implementare i meccanismi basilari astratti in forma teorica in dispositivi meccanici – principalmente elettronici – che funzionassero allo stesso modo, replicando di fatto il comportamento o la funzione dell'oggetto osservato.

La scommessa epistemologica della cibernetica consiste esattamente nel pensare che, una volta che i meccanismi fossero stati replicati in un artefatto artificiale, allora sarebbe stata dimostrata la validità della teoria che ne aveva ispirato la costruzione. Il celeberrimo assunto, che verrà poi formalizzato come principio cardine del funzionalismo, si costituisce come una linea guida non solamente per la ricerca del funzionamento della mente e del pensiero, ma soprattutto come un'ulteriore e novecentesca realizzazione dell'ambizione dell'umano. Infatti, è proprio qui che è possibile individuare il momento in cui la robotica come disciplina trova la sua condizione nascente di possibilità e acquisisce lo scopo che ha ancora oggi, ossia quello di realizzare artefatti artificiali che *simulino* gli esseri umani e che si dimostrino capaci di svolgere alcune delle funzioni che caratterizzano questi ultimi essenzialmente come tali (Dupuy 1994; Boden 2006).

In questa operazione, il metodo sintetico<sup>31</sup> si trova ad essere un elemento fondamentale da tenere in considerazione per comprendere al meglio come si sia evoluta la concezione di *replicabilità* e di *simulazione*, e come il processo di verifica delle teorie elaborate potesse passare per una sua realizzazione pratica, contribuendo quindi sostanzialmente allo sviluppo della mente come macchina. In generale, l'attività del metodo sintetico è quella di «build[ing] up theories of a system by attempting to construct an artificial system that exhibits the same capabilities as the natural system» (Steels 1995: 90), procedendo al controllo delle teorie elaborate tramite un confronto (*test*) tra il comportamento del sistema osservato, descritto dalle teorie ipotizzate, e il sistema costruito per implementarle (Ibid.: 91). In questo tentativo di comprensione tramite la costruzione (*understanding by building*, Langton 1989), il concetto di macchina che viene convocato non riguarda solamente l'istanziatura fisica, la congerie di meccanismi fisici ed elettronici, ma un livello superiore di astrazione, di riproduzione, che arriva a identificarsi con la struttura astratta del comportamento comune a sistemi naturali,

---

<sup>31</sup> Lo sviluppo del metodo sintetico è ben documentato in letteratura e trova una delle sue contemporanee declinazioni negli studi sulla *Artificial Life* (AL). Per una ricostruzione storica del metodo sintetico Cordeschi (1998, 2008a, 2008b); per una relazione sull'attuale stato dell'arte della disciplina dell'AL (Aguilar et al. 2014).



oggetto dell'osservazione, e sistemi artificiali costruiti per verificarne la validità. Come sottolineato da Cordeschi (2008a: 167), il metodo sintetico

non ha mai avuto a che fare con la proposizione di più o meno riuscite *metafore* sulla cui base confrontare organismi viventi e macchine, ma ha sempre avuto a che fare con *modelli*, intesi come artefatti fisici funzionanti di diversa natura, i quali possono contribuire alla spiegazione di alcune caratteristiche essenziali degli organismi viventi (l'adattamento, l'apprendimento, lo sviluppo, e l'evoluzione, ecc.).

In sostanza, dunque, nel metodo sintetico assume una posizione centrale l'artefatto, non semplicemente come metafora dell'organismo attraverso l'osservazione della quale è possibile capire qualcosa sull'oggetto di cui il sistema artificiale è metafora, ma come mezzo di controllo *funzionante* di una teoria elaborata riguardo al suo comportamento. Questa capacità, ovvero quella delle macchine (anche fisiche) di mettere in scena un comportamento in un ambiente, garantita dalle innovazioni tecnologiche soprattutto in ambito elettronico, rendeva necessario un allargamento del concetto di macchina, che passava quindi dall'essere un congegno meccanicistico il cui movimento era giustificato da rigidi automatismi (come quelli dell'epoca appena precedente), a una macchina capace di registrare stimoli provenienti dall'esterno e in qualche modo reagirvi, elaborando di fatto quello che veniva descritto come un *comportamento*. Tale salto di qualità del concetto di macchina viene esplicitamente sottolineato da uno degli ossimori che definiscono la cibernetica e che si configura come la condensazione di un programma di ricerca nella sua interezza: quello di studio della macchina vivente, o macchina teleologica<sup>32</sup> (Dupuy 1994: 34).

A partire da una serie di riconsiderazioni del concetto di macchina, prendono via alcune riflessioni sul ruolo che questa, in quanto *modello* della teoria, assume nel processo di spiegazione del fenomeno interessato. Pioniere di questa riflessione è quello che è stato definito "l'approccio del robot" di Thomas Ross, forse la prima ufficiale e dichiarata applicazione del metodo sintetico (Cordeschi 1998). L'obiettivo di Ross era non solo quello di realizzare esemplari di macchine artificiali che funzionassero secondo i principi elaborati nella teoria psicologica, fornendone quindi una verifica empirica, ma anche di poter controllare le diverse ipotesi psicologiche sulla natura del pensiero. Il punto era costruire un percorso epistemologico, una *epistemologia sperimentale* per come l'avrebbe definita poi McCulloch (1965), in cui, evitando il riduzionismo, al processo di messa in forma artificiale della teoria in un artefatto funzionante seguiva una riflessione sulla base della comparazione del funzionamento della macchina con i principi teorici che l'avevano prodotta (Rosenblueth & Wiener 1945). In questo modo, la macchina stava condensando due sensi in cui il suo ruolo di *modello* poteva essere inteso: da una parte, il modello è qualcosa che astrae il

---

<sup>32</sup> Il movimento teleologico è un altro dei concetti centrali sviluppati dalla cibernetica. Per gli sviluppi sulla teleologia nel movimento cibernetico: Rosenblueth et al. (1943); Wiener (1948, 1950); von Bertalanffy (1951).

funzionamento dell'oggetto di riferimento, consentendo un maggior grado di manipolabilità affinché si affini la comprensione dell'oggetto stesso; dall'altra, il modello è qualcosa che deve essere preso ad esempio, posto al centro della considerazione scientifica affinché possa guidare lo sviluppo dell'impresa di conoscenza<sup>33</sup>. La vera innovazione al cuore del metodo sintetico è dunque il rapporto tra teoria e modello, che si trova a costruirsi in un modo del tutto particolare rispetto al metodo scientifico classicamente inteso, quest'ultimo basato sulla dimostrazione matematica. Nel metodo sintetico, la macchina si trova ad essere sia il termine di un ragionamento deduttivo, che va dalla teoria alla sua realizzazione artificiale, sia l'incipit di un processo di inferenza induttivo, che fa ritornare sulla forma della teoria attraverso l'osservazione delle particolarità del comportamento istanziate dal modello artificiale. Prende forma un modello elicoidale della spiegazione (Cordeschi 2008a; 2004), che si modifica rispetto ai modelli classici grazie al fatto che il modello, la macchina, può realizzare conseguenze del modello, la teoria, nel suo comportamento che nell'ipotesi teorica primigenia potevano non essere dappprincipio previste. Il ritorno sulla teoria, con il fine di completare il piano astratto con le nuove caratteristiche individuate tramite il primo passaggio del metodo, è garantito dalla equivalenza funzionale posta in essere dal metodo stesso. Ciò che conta, dunque, per mettere a paragone il funzionamento di entrambe le istanze osservate, la teoria e il modello, è il livello dell'organizzazione, non quello della realizzazione fisica (Tamburini e Datteri 2005).

Sul ruolo del modello nel metodo sintetico verranno poi sviluppate numerose riflessioni, soprattutto in merito alle restrizioni da applicare al modello empirico affinché possa effettivamente spiegare i principi della teoria che istanzia (Cordeschi 2008a; Datteri 2010). Il problema del modello, d'altronde, era stato introdotto già in una delle pubblicazioni avvenute nell'anno *mirabilis* della cibernetica, il 1943, in cui Rosenblueth, Wiener e Bigelow dichiaravano che «The ultimate model of a cat is of course another cat, whether it be born of still another cat or synthesized in a laboratory.» (Rosenblueth et al. 1943: 23). Risulta di particolare interesse qui la considerazione dei pionieri della cibernetica, perché è proprio alla costruzione di un robot che fanno riferimento per esplicitare il loro ragionamento sul ruolo che la costruzione di un artefatto possa avere nella pratica scientifica di comprensione del comportamento:

If an engineer were to design a robot, roughly similar in behavior to an animal organism, he would not attempt at present to make it out of proteins and other colloids. He would probably build it out of metallic parts, some dielectrics and many vacuum tubes. The

---

<sup>33</sup> «Un modello è una forma astratta, per così dire, che si incarna o è istanziato dai fenomeni. Molto diversi domini della realtà fenomenica [...] possono essere rappresentati da modelli identici, che stabiliscono una relazione di equivalenza tra di loro. Un modello è la classe di equivalenza corrispondente. Si gode quindi di una posizione trascendente, non dissimile da quello di una Idea platonica di cui la realtà è solo una pallida imitazione. Ma il modello scientifico è artificiale. È a questo punto che la relazione gerarchica tra l'imitatore e imitato viene ad essere invertita. Anche se il modello scientifico è una imitazione della natura umana, lo scienziato è incline a considerarla come un "modello", nel senso comune, di natura. Così la natura è ritenuta imitare il modello con cui l'uomo cerca di imitarla» (Dupuy 1994: 64)

movements of the robot could readily be much faster and more powerful than those of the original organism. Learning and memory, however, would be quite rudimentary. In future years, as the knowledge of colloids and proteins increases, future engineers may attempt to design the robot not only with behavior, but also with a structure similar to that of a mammal. (Rosenblueth et al. 1943: 23)

Tale argomentazione sembra quasi una profezia se messa in relazione a quello che sarebbe avvenuto pochi anni dopo, con la costituzione della robotica e i primi tentativi di realizzare robot a partire non da materiale organico, ma da parti metalliche; non avevano però previsto l'arrivo dell'intelligenza artificiale, di cui comunque questo articolo è considerato uno dei pilastri fondativi.

L'articolo del '43 di Rosenblueth, Wiener e Bigelow non è l'unico elemento in comune tra la cibernetica e l'approccio dell'intelligenza artificiale. Per entrambi, infatti, a prescindere dalle restrizioni che avvicinano più o meno il modello alla funzione modellizzata, ciò che garantisce il rapporto di verifica e spiegazione tra l'organismo di cui la teoria ipotizza un meccanismo di funzionamento e l'artefatto artificiale che lo incarna è il "principio della realizzabilità multipla", fondamentale non solamente nella prima cibernetica e nel metodo sintetico, ma anche per l'influenza che ha avuto nelle tesi formulate dal funzionalismo computazionalista. Questo principio garantisce all'indagine sintetica la dignità scientifica, permettendogli di stagliarsi tra le altre metodologie di indagine della mente per la sua efficacia pragmatica (data l'immediatezza della possibilità di realizzazione e di verifica dei principi guida) e per l'aderenza alle esigenze del tempo (Cordeschi 1998, 2004, 2008b). Questo principio non solo caratterizzerà la cibernetica, ma resterà un pilastro fondamentale nelle successive indagini sui processi mentali, attraversando indistintamente diverse correnti di pensiero, sia nell'approccio simbolico sia in quello sub-simbolico, e costituirà un minimo comun denominatore anche per gli sviluppi più recenti della concezione della mente come artefatto, come ad esempio la *behavior-based robotics* (Brooks 1991a; cfr. Cordeschi 2008a; §2.3.2).

In sostanza, ciò che rende possibile, insieme a una serie complessa di circostanze concomitanti, la possibilità di concepire la meccanizzazione della mente è la comparsa di un nuovo tipo di macchina: la *macchina cibernetica*. Questo nuovo tipo di macchine rendeva possibile incorporare concetti sviluppati all'interno della teoria cibernetica, come quello di *feedback* (positivo e negativo)<sup>34</sup>, di omeostasi<sup>35</sup>, di servomeccanismo<sup>36</sup>, di auto-organizzazione<sup>37</sup>, o di fenomeni studiati su organismi biologici come il comportamento, l'apprendimento, o modelli neurali. In più, la comparsa di questo

---

<sup>34</sup> «[...] quando intendiamo effettuare un movimento secondo un determinato modello, la differenza fra tale modello e l'effettivo svolgersi del movimento viene adoperata come un nuovo segnale che determina una regolazione del movimento stesso tale da mantenerlo quanto più possibile vicino a quello dato dal modello» (Wiener 1948: 29).

<sup>35</sup> «Un vasto insieme di casi in cui qualche tipo di feedback non solo è individuabile nei fenomeni fisiologici ma è assolutamente essenziale alla continuazione della vita rientra in ciò che è noto come omeostasi» (Wiener 1948: 156).

<sup>36</sup> Il servomeccanismo si configura come un insieme di feedback (cfr. Wiener 1948).

<sup>37</sup> «“self-organizing,” i.e., that [a system] can be determinate and yet able to undergo spontaneous changes of internal organization» (Ashby 1947: 125).

nuovo tipo di macchina consente alla metafora dell'umano come macchina di non esaurire la sua capacità euristica al livello della modellizzazione del corpo, ma scalare e arrivare a costituire uno strumento anche per l'indagine della mente. L'indagine di Cordeschi (1998) ci mostra che, anche se questo tipo di macchine sono diventate celebri con il loro impiego nella cibernetica, è interessante notare come alcuni esemplari fossero stati costruiti già tra la fine dell'Ottocento e l'inizio del Novecento. Nel caso specifico, il primo esempio di tale macchina è il "cane elettrico" Selenio, inventato e costruito da Hammond e Miessner intorno alla metà degli anni Dieci del '900. Selenio, anche se non ricordava esattamente le fattezze di un cane, era capace di muoversi in un ambiente controllato in cui fossero presenti delle fonti luminose. La macchina, infatti, incorporava dei meccanismi eliotropici esplicitamente teorizzati per il genere animale delle falene da Jacques Loeb, e fu riconosciuta dallo scienziato stesso come realizzazione pratica e conferma della sua teoria. La macchina eliotropica riusciva a seguire una fonte luminosa o a sfuggirvi, senza però incorporare ancora alcun meccanismo di apprendimento (cfr. Cordeschi 1998: 12-22). Ciò sarebbe stato reso possibile per la prima volta con la macchina idraulica di Bent Russel, che tentava di realizzare i principi sviluppati riguardo l'apprendimento neuronale da Spencer, Meyer e Thorndike, simulando alcuni elementi essenziali della plasticità delle connessioni nervose (Cordeschi 1998: 81-90).

Senza dubbio più sofisticate, ma aventi lo stesso ruolo di conferma e sviluppo della teoria che incorporavano, sono le macchine cibernetiche vere e proprie, costruite tra gli anni Quaranta e Cinquanta. Non è possibile non citare le più celebri, come il topo elettronico ideato e costruito da Claude Shannon, che contribuirà all'impresa di conoscenza cibernetica soprattutto con la teoria dell'informazione. Presentata alla ottava delle *Cybernetic* (poi diventate *Macy's*) *Conferences* nel 1952, la macchina era soprannominata Teseo ed era una riproduzione in forma elettromeccanica a grandezza reale di un topolino capace di attraversare un labirinto articolato su un quadrato 5x5, imparando il percorso migliore per uscirvi tentativo dopo tentativo. Due motori elettrici permettevano al topolino meccanico di muoversi nel labirinto, registrando in memoria gli "errori" quando urtava una delle pareti, ma azzerandosi quando il labirinto non veniva risolto in un certo numero di mosse, permettendo di scappare dal circolo vizioso delle ripetizioni (Boden 2006: 221). Nonostante la semplicità dell'algoritmo che incorporava e della dotazione elettronica di cui Shannon si avvalse, il topolino Teseo suscitò la meraviglia di molti dei presenti e diventò un'istituzione per la robotica che doveva ancora formarsi<sup>38</sup> (Polson e Scott 2018: 70-71, ad es. Moravec 1988).

Insieme a Teseo, le tartarughe di William Gray Walter (1950a, 1950b, 1951; 1963) si guadagnano un posto d'onore tra le macchine cibernetiche, costituendone forse l'esempio più emblematico e la

---

<sup>38</sup> Anche se ricevette forti critiche da alcuni degli scienziati presenti alla conferenza, i quali sottolineavano il fatto che il topo meccanico non cercasse di riprodurre alcuna funzione presente nel topo biologico (cfr. Dupuy 1994).

prima vera realizzazione di un robot autonomo. Anche in questo caso, l'assimilazione della macchina a una forma animale è suggerita più dall'immaginazione dei costruttori che dall'estetica delle macchine stesse. Nel caso delle macchine di Gray Walter, queste potevano ricordare le tartarughe perché erano coperte da una sorta di corazza e perché la velocità non era esattamente una loro caratteristica identificativa. Tra queste, senza dubbio la più celebre è la *Machina speculatrix*, una classe di automi costruita tra il 1948 e il 1949 per dimostrare che connessioni tra un numero limitato di cellule cerebrali potessero dare luogo a comportamenti complessi, ipotizzando dunque che il segreto del funzionamento del cervello risiedesse nelle modalità in cui era cablato e non necessariamente nella quantità delle cellule che lo componevano. Influenzato anche dalle idee di Pavlov sul condizionamento, l'intenzione di Grey Walter era di modellare il comportamento di "goal seeking" e, in seguito, anche l'apprendimento sfruttando il minor numero di meccanismi possibili e un modello molto semplice di sistema nervoso. Le sue tartarughe avevano solamente due valvole, due relè, due motori, due condensatori e un sensore (che poteva essere tattile o fotosensibile): in sostanza, solamente due neuroni (Boden 2006: 224). Ma riuscivano a infondere un senso di vitalità nell'osservatore: la prima *Machina speculatrix*, ribattezzata ELSIE (Electro-mechanical robot, Light-Sentitive with Internal and External stability), esplorava il suo ambiente, usando la sua cellula fotoelettrica per cercare la luce, ammirarla se adeguatamente calibrata, allontanarsi se troppo forte e avvicinarla se ancora troppo debole. Grey Walter parla di ELSIE in termini deliberatamente biologici, descrivendola come una tartaruga in grado di mostrare un tropismo positivo a una luce moderata, ma negativo nei confronti di una luce troppo forte o in condizioni di buio. Il lavoro di Grey Walter, non ridotto solamente alla costruzione delle tartarughe fototropiche ma sviluppato anche agli esemplari successivi, fu essenziale per lo sviluppo di una robotica *biologically based*, e verrà riconosciuto in seguito da tutta una generazione di robotici quali Micheal Arbib, Valentino Braitenberg, Rodney Brooks, Maja Mataric, Randall Beer e Barbara Webb come primo esempio della possibilità di costruire macchine dal comportamento complesso imitando i processi biologici degli organismi viventi.

Ultimo esempio che è necessario citare per definire le macchine cibernetiche è l'omeostato di William Ross Ashby (1952). Costruito per illustrare e testare il suo concetto di ultrastabilità, l'omeostato si basa sull'idea di un sistema regolatore omeostatico che mantiene un equilibrio dinamico interno all'organismo attraverso il controllo e il monitoraggio dei suoi parametri interni ed esterni (Ibid: 163-188). In altre parole, lo scopo della macchina omeostatica era incorporare il concetto di *feedback negativo* come meccanismo di mantenimento dell'ordine del sistema: il principio si traduceva in una macchina che mostrava come sia i sistemi biologici sia i sistemi artificiali potessero mantenere una stabilità e adattabilità a fronte di perturbazioni esterne che vengono assorbite

attraverso una compensazione, detta appunto omeostatica. Il focus era, come previsto dal metodo cibernetico, sullo sguardo dell'osservatore e sulla capacità simulativa della macchina: come sottolineato da Fabbris nell'introduzione all'ultima edizione italiana di *Design for a brain* (Ashby 1952), «nell'idea di Ashby Homeostat doveva testimoniare di come un comportamento che a un osservatore può apparire adattivo e finalizzato non sia nient'altro che l'esito di un processo di variazione casuale e selezione» (Fabbris 2021: 34).

Sicuramente, l'interesse suscitato dall'omeostato non era della stessa meravigliata natura di quello rivolto alle tartarughe di Grey Walter, tanto che quest'ultimo non esitò a definirlo *Machina sopora* (ovvero macchina che dorme), né tanto meno lo scopo dell'omeostato era quello di replicare comportamenti "superficiali". Invece, lo scopo di Ashby era quello di esplorare le caratteristiche più generali della vita e della mente, ovvero il concetto di feedback e quello di adattività auto-organizzata (Boden 2006: 228). Anche se all'inizio suscitò qualche scetticismo, questa macchina cibernetica assumerà uno dei ruoli più importanti per la formulazione e la dimostrazione dei principi cibernetici, influenzando numerose ricerche in robotica e in *Artificial Life (A-Life)* (Holland 1997; Langton 1997).

Guardando a questi esempi emblematici, possiamo allora delineare le caratteristiche generali di quella che abbiamo definito la *macchina cibernetica*. In primo luogo, è una macchina elettronica, perché costruita a seguito di – e grazie a – una importante serie di innovazioni in ambito tecnico, ovvero l'avvento dei componenti attivi, come le valvole termoioniche, i semiconduttori, i transistori e i circuiti integrati. Ciò «permise per la prima volta di distinguere tra il *segnale*, consistente in bassa corrente o in informazione e così via, e l'azione, rappresentata da alta corrente o energia o moto e così via» (Cordeschi 1998: 213). In secondo luogo, la macchina cibernetica si distingue perché viene costruita per testare una teoria, un'idea proveniente da studi sul funzionamento del cervello o circa il comportamento delle forme di vita biologiche, dando così l'avvio alla teoria connessionista della mente (Cordeschi 1998) e costituendosi di fatto come una delle prime attestazioni di quello che verrà poi chiamato il metodo sintetico (Cordeschi 2008a, 2008b).

Questo ci conduce alla terza caratteristica da sottolineare, legata alla linea direttrice dell'ideologia cibernetica e che sorregge l'impalcatura teorica e operativa: la *macchina* non è più solamente considerata una realizzazione pratica, un insieme di meccanismi, circuiti, ingranaggi, che assemblati e fatti funzionare producono un qualche tipo di comportamento, ma uno strumento teorico, astratto, in grado di incarnare sia il funzionamento della vita, sia quello della macchina empirica. Quello che possiamo definire come il movimento scalare, che porta la macchina non più a denotare solamente artefatti modellizzati meccanicisticamente ma a funzionare da concetto chiave per unificare il mondo della vita con quello delle realizzazioni empiriche artificiali, è meravigliosamente descritto da Morin (1977: 193-194):

Mentre i motori giocano con il fuoco, le macchine automatizzate giocano alla vita. A partire dai meccanismi e dai dispositivi a orologeria (XIII secolo) si è sviluppato un automatismo di operazioni sempre più precise, delicate e diversificate, che costituiscono catene operanti ad anello su se stesse in modo reiterativo. Si è così arrivati agli automi del XVIII secolo, che imitano in modo stupefacente i gesti del comportamento animale e umano. Questo automatismo orologistico si è sviluppato nei meccanismi industriali, fino all'apparizione di uno stadio nuovo di complessità nell'automatismo macchinale: lo stadio cibernetico. A questo punto, un comando sino a quel momento esterno diviene interno (programma) e organizzatore (computer) e l'automa cibernetico si mette a somigliare al vivente non più nei tratti esteriori, come l'automa orologistico, ma per l'organizzazione del comportamento.

Con la cibernetica, dunque, ribadiamo, si attua quella saldatura che conduce all'identificazione della mente come macchina, ma non una macchina "orologistica", quanto invece una macchina intesa come principio di funzionamento del comportamento. Le nozioni elaborate nel contesto cibernetico, come quella di feedback positivo e negativo, di ultrastabilità, di auto-organizzazione si riferiscono al funzionamento di una macchina generale, quella che riesce ad abbracciare sotto di sé tutte le sue espressioni: gli organismi, il comportamento, la struttura cerebrale, i meccanismi di retroazione, la vita stessa sono istanziazioni della macchina come strumento teorico *universale*.

È in questo frangente che si può iniziare a parlare di meccanizzazione della mente, ma non solo perché è uno dei propositi principali della cibernetica. Contemporaneamente, infatti, una distinta linea di ricerca si stava delineando e imboccava un binario parallelo, con lo stesso scopo di meccanizzare il funzionamento della mente ma portata avanti da personalità differenti e orientata allo studio dei meccanismi logici del pensiero prettamente umano, in alternativa all'osservazione del mondo vivente e del suo comportamento generale. La linea di ricerca in questione, definita simbolico-computazionale (Davis 2000; Boden 2006), è ovviamente quella che ha dato vita al progetto dell'Intelligenza Artificiale, che nasce ufficialmente nel 1956 alla conferenza di Dartmouth ma che vede stabilite le sue condizioni di possibilità nell'invenzione di un altro tipo di macchina: quella computazionale.

#### 1.2.4 *Macchine computazionali*

In sostanza, entro la fine degli anni '40 del '900, due erano le principali modalità di indagare la mente in quanto macchina: una era quella cibernetico-dinamica, che abbiamo brevemente riassunto nel paragrafo precedente, l'altra quella simbolico-computazionale. L'ipotesi conoscitiva centrale dell'approccio simbolico-computazionale consiste nel pensare che la mente possa essere compresa e simulata attraverso la computazione simbolica dei dati. In questo approccio la mente è considerata come un sistema di elaborazione delle informazioni, in cui i simboli rappresentano i concetti, e le regole di manipolazione dei simboli sono l'equivalente dei processi mentali. In sostanza, la mente viene trattata come una macchina che elabora simboli in modo sequenziale dando come frutto

l'intelligenza attraverso la manipolazione formale di tali simboli (Boden 2006: 169-182). Tale approccio si differenzia da quello cibernetico per diverse ragioni (cfr. Boden 2006: 232-233):

- i) Le radici storiche erano profondamente differenti: la cibernetica era costruita sull'ingegneria del controllo, sulla fisiologia generale, la riflessologia, e la teoria dell'informazione; l'approccio simbolico si fondava sulla logica proposizionale e la filosofia del linguaggio logicista.
- ii) L'oggetto di indagine (i meccanismi di controllo da una parte, la struttura del pensiero logico dall'altra) costituisce una delle differenze maggiori: da una parte, la cibernetica si concentra sull'imitazione del cervello come macchina che produce il comportamento; dall'altra, l'approccio simbolico-computazionale guarda ai meccanismi logici del pensiero e alla struttura astratta del ragionamento.
- iii) Allo stesso tempo, tale differenza di visioni e la presa in carico di diversi oggetti di indagine si concretizza in un approccio organicista della cibernetica, interessata a guardare i fenomeni nel loro intero e a riprodurre il comportamento complesso dell'organismo nel modo più accurato possibile, e in un approccio "spacchettante" dei computazionalisti, impegnati nell'analisi dei problemi a partire dalla loro scomposizione in sotto-problemi più piccoli. Ciò comporta anche l'adozione di approcci differenti, come nel caso della matematica (la cibernetica adotta un approccio continuo-statistico, l'approccio computazionale uno discreto).
- iv) Le differenze rispetto all'impianto teorico ed epistemologico si traducono nella costruzione di macchine differenti. Infatti, i risultati tecnologici per la cibernetica hanno condotto all'ingegneria del controllo e alle forme dinamiche del connessionismo, mentre per l'approccio computazionale hanno portato alla Intelligenza Artificiale simbolica. La differenza sotto questo aspetto è ancor più apprezzabile se si guarda alle macchine effettivamente realizzate da entrambi gli approcci: ad esempio, gli artefatti artificiali enumerati nel precedente paragrafo portano quasi tutti nomi di animali, mentre nel caso dell'approccio simbolico non si ricorrerà ad alcun parallelo con il mondo degli organismi viventi.

Le differenze tra i due approcci, forse fin troppo acute nelle ricostruzioni storico-filosofiche fino a restituire un rapporto di inconciliabilità fra gli stessi, sono tuttavia bilanciate da una serie di tratti comuni. Innanzitutto, l'obiettivo delle indagini era fondamentalmente lo stesso: la comprensione della mente e dei processi mentali. Entrambi gli approcci, di marcato stampo interdisciplinare, cercano di sviluppare modelli e teorie che spieghino come la mente funzioni, congiuntamente al



tentativo di replicarla o simularla artificialmente. Per raggiungere tale scopo, entrambi gli approcci si affidavano a un modello computazionale della mente: se nella cibernetica questo si traduceva con l'elaborazione dei concetti di feedback, controllo e auto-organizzazione, nell'approccio simbolico-computazionale la chiave era l'algoritmo<sup>39</sup>. Entrambi gli approcci considerano la mente come un sistema che elabora e manipola informazione. Infine, come abbiamo precedentemente accennato, il principio di equivalenza tra le funzioni, svolte indistintamente da un essere vivente o da una macchina, costituisce un altro punto in comune tra i due approcci, anche se, come vedremo, declinato in modi diversi. È esattamente dalla comunanza del principio dell'equivalenza delle funzioni che si vuole partire per mostrare come anche la *macchina computazionale* sia un'ulteriore realizzazione dell'ambizione dell'umano, passata attraverso la metafora dell'umano come macchina e sviluppata fino a impiegare la macchina per conoscere il funzionamento della propria mente. Per la nostra indagine sulla costituzione della robotica, inoltre, questo passaggio risulta cruciale, in quanto è proprio questo il tipo di macchina che verrà adottata fin dalla costruzione dei primi robot.

È fondamentale, allora, prendere in considerazione le caratteristiche principali di questa macchina computazionale, che faremo meglio a chiamare con il suo nome originale: la macchina di Turing. Alan Turing, considerato uno dei padri fondatori dell'intelligenza artificiale, sviluppa il modello della sua macchina nel tentativo di risolvere il cosiddetto *Entscheidungsproblem*, ovvero il problema della decisione elaborato dal matematico David Hilbert nel 1928 che riguarda la possibilità di determinare se una data affermazione matematica possa essere dimostrata vera o falsa tramite un processo algoritmico (Davis 2000: 190-192). Con la costruzione di questa macchina, Turing introduceva anche una definizione formale del concetto di *computabilità*, ponendo quindi le condizioni formali che definissero cosa significhi che un problema può essere risolto da un processo algoritmico (Turing 1936). Il sostanziale avanzamento nella dimostrazione del problema della decisione fu fatto nel dimostrare<sup>40</sup> che ogni algoritmo computabile potesse essere computato da una macchina di Turing, e che questo non era possibile per il primo ordine della logica predicativa: in breve, non esiste un algoritmo generale in grado di risolvere il problema per tutte le possibili proposizioni matematiche (Davis 2000: 182-218). Ma come era fatta questa macchina?

Nello stesso articolo del 1936, Turing descrive la sua macchina come un *immaginario* dispositivo meccanico in grado di eseguire analiticamente qualsiasi operazione di calcolo che fosse possibile portare a termine da un essere umano. La macchina è composta semplicemente di una testina di lettura/scrittura e di un nastro (infinito) diviso in celle. La testina può leggere o scrivere simboli sul

---

<sup>39</sup> Gli algoritmi sono insieme di istruzioni che possono essere computate per raggiungere un risultato o un obiettivo. (Moschovakis 2001)

<sup>40</sup> Nonostante tale visione sia entrata nel senso comune e al giorno d'oggi costituisca la versione più diffusa e accettata del teorema di Turing per la computabilità, la questione sul fatto che quella di Turing sia un'ipotesi più che una dimostrazione è tutt'oggi discussa (cfr. Davis 2000; Cappuccio 2005; Boden 2006: 179).

nastro e spostarsi in avanti o indietro di una cella alla volta a seconda delle istruzioni date alla macchina (Turing 1936: 231-232). L'azione che la macchina fa nel leggere e/o scrivere il simbolo sul nastro era esattamente quello a cui Turing si riferiva come computazione. Tanto semplice quanto geniale, il paragone istituito da Turing era con l'attività umana osservata nella a lui contemporanea figura professionale del *computer*, ovvero per la maggior parte delle volte donne la cui mansione era quella di svolgere calcoli manualmente. Il concetto di algoritmo veniva dunque rintracciato tanto nella macchina, quanto nello svolgimento della funzione da parte di un calcolatore umano. Come sottolineato da Davis (2000: 192),

Turing sapeva che un *algoritmo* è tipicamente definito un elenco di regole che una *persona* può eseguire in modo meccanico e preciso, come le ricette dei libri di cucina, ma spostò la sua attenzione dalle regole a quello che la persona effettivamente *faceva*, ed eliminando uno dopo l'altro i dettagli inessenziali riuscì a dimostrare che una tale persona poteva limitarsi a poche azioni di base, estremamente semplici, senza che il risultato finale del calcolo cambiasse. Poi, facendo un altro passo, comprese che l'essere umano poteva essere sostituito da una macchina capace di eseguire quelle stesse azioni di base; e infine dimostrò che nessuna macchina in grado di eseguire tali azioni poteva stabilire se una data conclusione era derivabile dalle premesse date [...], e concluse che non esisteva un algoritmo per l'*Entscheidungsproblem*. (*corsivi dell'autore*).

In questa ricostruzione, due sono le osservazioni che è opportuno riportare: da una parte, il focus di Turing viene rivolto alla *funzione*, cioè al meccanismo attraverso cui è possibile realizzare lo stesso processo nella mente umana e nella macchina; dall'altra, la funzione deve essere spogliata dei suoi elementi addizionali, ovvero fatti e caratteristiche contingenti allo svolgimento dell'atto nel calcolatore umano che non richiedono di essere inclusi nella descrizione astratta. Non fa differenza nel processo computazionale, infatti, se il calcolatore umano beve un bicchiere di succo di frutta mentre svolge il suo compito, né se ha le gambe accavallate o lo smalto alle unghie: ciò che è fondamentale è riuscire ad astrarre tutte le componenti che nel processo di computazione fanno la differenza, escludendo tutto ciò che non genera direttamente il risultato voluto (ovvero il risultato esatto del processo di computazione). In sostanza, lo scopo era quello di costruire il giusto algoritmo. Ciò risulta importante perché fornisce un ulteriore sviluppo per l'implementazione della mente nella macchina, la quale deve epurarsi dalle contingenze della computazione mentale: il supporto attraverso cui viene svolta la funzione perde in questa sede qualsiasi rilevanza. Il principio dell'equivalenza delle funzioni viene qui portato alla sua massima astrazione, dematerializzando qualsiasi meccanismo nella sua sequenza logica. Il risultato è quindi quello di avere in sostanza due macchine: quella corporea, dominio della biologia e della chimica, che può ancora rispondere a delle leggi fisiche che governano la materia; e quella mentale, governata invece da meccanismi astratti del pensiero, la cui struttura è individuabile tramite strumenti logici e leggi sintattiche di formazione linguistica. Senza dubbio, per la corrente simbolico-computazionale il focus era sulla macchina mentale, completamente

diversa da quella fisica che si pensava non avesse alcuna influenza sulla prima, la quale quindi assumeva un'autonomia di indagine ancora maggiore rispetto ai tentativi precedenti<sup>41</sup>.

Le innovazioni apportate da Turing in questa direzione non si esauriscono con l'ideazione della macchina di Turing standard. Una ulteriore versione risulta ancora più utile e di interesse per capire le implicazioni delle idee di Turing per la robotica: la *macchina universale*. La macchina di Turing standard era immaginata poter risolvere solo un problema computazionale alla volta, e ha una definizione precisa di stati, simboli e regole di transizione. Ciò implica che, una volta programmata per eseguire un certo algoritmo, una macchina di Turing standard può eseguire solo quel particolare algoritmo e non può essere facilmente adattata per computarne di altri. La macchina universale di Turing, invece, sorpassa queste difficoltà: si tratta infatti di un tipo speciale di macchina di Turing, in grado di simulare qualsiasi algoritmo computazionale. In sostanza, la sua universalità è data dal fatto che è capace di leggere una descrizione di un'altra macchina di Turing standard insieme ai suoi dati input ed eseguire l'algoritmo descritto da quest'ultima. Ciò significa che una macchina di Turing universale può essere programmata per eseguire qualsiasi funzione, a patto che sia esplicitato ogni passaggio, rendendola estremamente versatile e potente e costituendo una delle prime possibilità di implementare la computazione in una macchina *fisica* (Boden 2006: 172-173). Ciò che risulta di fondamentale importanza riguardo la macchina di Turing universale è che mette in gioco la nuova nozione di *programmabilità*: dal momento che questa può eseguire qualsiasi algoritmo computabile, il comportamento della macchina può essere modificato dando nuove istruzioni sulla banda di memoria e consentendo una vastità di impieghi senza precedenti, tutti condensati in un'unica realizzazione macchinica.

Grazie a queste due nozioni, quella di funzione nel suo senso più astratto che si formalizza in algoritmo e quello di macchina di Turing universale con la sua programmabilità, si configura quella differenza sostanziale tra *software* e *hardware* che resterà fondamentale in informatica e nelle scienze cognitive fino ai giorni nostri. In generale, con *hardware* si fa riferimento alle componenti fisiche di un sistema informatico, compresi i processori, la scheda madre, la memoria, i dispositivi di archiviazione e le periferiche di input/output; il *software*, invece, consiste nelle istruzioni digitali che vengono eseguite analogicamente da questi componenti hardware. Nel caso della macchina di Turing fisica, ovvero realizzata e non immaginaria<sup>42</sup>, l'hardware consisteva nei nastri per l'archiviazione dei

---

<sup>41</sup> Come sottolineato da Ceruti, nelle pagine dedicate alla ricostruzione del ruolo della macchina di Turing per una teoria della cognizione: «Per definire una “macchina” non è assolutamente necessario riferirsi alle condizioni materiali che la realizzano o alle condizioni energetiche all'interno delle quali può funzionare. Ciò che conta è la regolarità del suo funzionamento. Più precisamente condizione necessaria e sufficiente per la definizione generale di macchina è che ogni stato della macchina sia univocamente definito dal suo stato interno precedente nonché dallo stato precedente dell'ambiente.» (Ceruti 1989: 58).

<sup>42</sup> Le prime versioni della macchina di Turing furono realizzate subito dopo la pubblicazione del suo articolo nel 1936. Tra le prime più celebri realizzazioni vanno annoverate l'Automatic Computing Engine (ACE), costruita da Turing in

dati, nelle testine per la lettura/scrittura dei simboli, lo stato della macchina, le regole di transizione e la struttura della macchina, mentre il software era composto dal programma (ovvero le istruzioni logiche che guidano il funzionamento della macchina), dei dati di input, intermedi e di output (Boden 2006: 175).

Questa divisione risulta cruciale non solo se vista nell'ottica della creazione del programma dell'intelligenza artificiale, ma anche per la linea di idee che sarebbe stata abbracciata nel periodo nascente delle scienze cognitive, le quali, congiuntamente, alla robotica avrebbero dato i natali. Infatti, la divisione tra hardware e software rendeva possibile rinforzare il principio dell'equivalenza funzionale incorporando esplicitamente la funzione (eseguibile attraverso il programma) dal supporto che la realizzava materialmente. Di conseguenza, in accordo allo stato della tecnologia dell'epoca e alle spinte date in questa direzione dal contesto bellico, la ricerca nei campi interdisciplinari coinvolti ha incanalato i suoi sforzi più che sulla realizzazione di creature che simulassero uno specifico comportamento organico più dettagliatamente possibile, verso la risoluzione di problemi inerenti la strategia, la logica e la programmazione di meccanismi utili alla risoluzione di problemi considerati di maggiore importanza e interesse. Come noto, questo porterà alla nascita nel 1956 a Dartmouth il progetto dell'Intelligenza Artificiale, progetto costruito da personalità afferenti principalmente alla corrente simbolico-computazionale come John McCarthy, Marvin Minsky, Allen Newell e Herbert Simon (Cordeschi e Frixione 2001), ma in cui convogliano contributi provenienti dalla cibernetica compatibili alla direzione già intrapresa<sup>43</sup>.

È interessante notare come, anche se questa celebre adunanza a Dartmouth era guidata dall'ipotesi programmatica che «every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it» (McCarthy et al. 1955 (2006): 12), la robotica non viene nemmeno citata come ambito di studio o come sfida da intraprendere: tutta l'attenzione era rivolta allo sviluppo di algoritmi, dei primi linguaggi di programmazione e di programmi veri e propri da implementare sui nuovi computer. In definitiva, il perdurare dell'attenzione sui progetti dell'intelligenza artificiale negli anni '50 e '60 del ventesimo secolo, preceduta e incentivata dalla costruzione dei primi grandi calcolatori come l'ENIAC (Electronic Numerical Integrator and Computer, 1943-1946) di John Mauchly e J. Presper Eckert o l'EDVAC (Electronic Discrete Variable Automatic Computer, 1949-1952) di John von Neumann, ha oscurato l'impresa di costruire organismi artificiali che imitassero la vita che aveva preso l'avvio nella ricerca cibernetica, posticipando ancora di qualche decennio la nascita della robotica.

---

Inghilterra durante la Seconda guerra mondiale, e la Turing Machine Mark I, una versione elettromeccanica della macchina di Turing che costituisce uno dei primi esempi di un computer a programma memorizzato (ovvero funzionante in base a un programma memorizzato su un supporto di memoria interno e non cablato) (Boden 2006: 177-187).

<sup>43</sup> Per una ampia panoramica sull'influenza della cibernetica per lo sviluppo dell'intelligenza artificiale si veda: Minsky 1961; Cordeschi 1994; 1997; 2004.

La macchina computazionale però, che fino a qui abbiamo visto costituirsi come uno strumento concettuale capace di far sviluppare riflessioni sulle modalità di realizzare l'intelligenza umana in artefatti elettronici, non esaurisce la sua spinta immaginifica all'interno delle discipline ingegneristiche. La portata dell'invenzione, infatti, trova fin da subito una profonda considerazione in ambito filosofico, proprio a partire da un altro articolo di Turing, pubblicato nel 1950. In questo articolo, l'inventore della macchina computazionale sonda per la prima volta le implicazioni sulla natura della mente e del pensiero che derivano dalla realizzazione, attraverso la nuova tecnologia delle macchine computazionali, di risultati considerabili intelligenti. «Can machines think?» (Turing 1950: 433) è la domanda che avrebbe fatto riversare i proverbiali fiumi di inchiostro alla quasi totalità dei filosofi della mente e scienziati cognitivi che avessero avuto lo scopo di affrontare il problema dell'intelligenza artificiale. L'articolo è passato alla storia non solo per la formulazione originaria di una nutrita serie di domande riguardo alla natura dell'intelligenza, su come riprodurla o sui necessari *caveat* derivati dall'impresa stessa enucleati in modo brillante da Turing, ma anche (e forse soprattutto) per la prima formulazione dell'«Imitation Game» (cfr. §5.1). Tale gioco, oggi celebre con il nome di Test di Turing, ha subito numerosi sviluppi e aggiustamenti (Copeland 2004; Longo 2009; Bianchini 2014) fino a diventare uno dei criteri accreditati per decretare se un artefatto costruito per simulare un'intelligenza umana abbia effettivamente raggiunto il suo scopo. In sintesi, la versione più celebre ad oggi diffusa del test coinvolge tre partecipanti: due utenti umani, uno dei quali necessariamente esperto, e una macchina. L'utente esperto interagisce con le altre due entità attraverso un sistema di comunicazione testuale, una chat potremmo dire oggi, senza che abbia visione diretta dell'entità con cui intrattiene la conversazione. Se, dopo una serie di scambi, l'utente esperto non riesce a capire chi tra i suoi due interlocutori sia la macchina e chi sia l'altro utente, allora si può dire che la macchina abbia superato il test: è stata scambiata per un essere umano.

Ponendo a margine gli straordinari contributi che tale formulazione ha dato per lo sviluppo dei meccanismi di intelligenza artificiale e per la concezione di intelligenza macchinica stessa, ciò che ci interessa in questa fase è che questo discorso comporta l'introduzione di un nuovo livello attraverso cui è possibile realizzare l'ambizione dell'umano. Infatti, se la simulazione delle funzioni per la riproduzione fedele del comportamento era stata la cifra distintiva della cibernetica, con la macchina computazionale Turing introduce una nuova caratteristica delle macchine: quella di *emulare* l'intelligenza umana. Questa divisione, che verrà poi ufficialmente formalizzata e resa celebre da Searle (1980) nella dicotomia *versione forte vs versione debole* dell'intelligenza artificiale<sup>44</sup>, si

---

<sup>44</sup> «According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right program can be literally said to *understand* and have other cognitive state». (Searle 1980: 417)

configura come un'ulteriore biforcazione delle traiettorie di ricerca sulla replicazione della mente. La strategia della *simulazione*, infatti, comporta non solo un'equivalenza tra le funzioni svolte, ma anche che l'obiettivo che guida la ricerca sia quello di tentare di istanziare gli stessi meccanismi che avvengono nella mente umana in una macchina artificiale, che riproduca fedelmente l'apparato funzionale: la metafora della mente come macchina si radicalizza, comportando un rapporto uno a uno tra mente e corpo da una parte e hardware e software dall'altra. D'altra parte, la strategia dell'*emulazione* comporta una equivalenza debole tra i meccanismi computazionali e il comportamento che ne consegue: ciò che risulta necessario e sufficiente per l'attribuzione dell'intelligenza a un artefatto artificiale è che questo dimostri un comportamento intelligente, senza bisogno alcuno che i meccanismi che generano tale comportamento siano equivalenti o addirittura uguali a quelli che danno luogo allo stesso comportamento nell'essere umano. Lo stesso Turing (1947) afferma che sarebbe impossibile per un essere umano raggiungere le capacità computazionali di un calcolatore, indebolendo quindi il paragone e arrivando a considerare i calcolatori come strumenti utili alla comprensione della mente, ma non perfettamente identificabili con questa. Tale ulteriore riflessione, introdotta da Turing ma seguita da diversi specialisti dell'intelligenza artificiale come von Neumann (cfr. Davis 2000; Boden 2006; Israel 2004), contribuisce largamente al processo di astrazione delle funzioni cognitive, oramai totalmente scorporate dal supporto che le istanzia e suscettibili di essere investigate anche attraverso la loro implementazione in programmi e software, a patto che possano realizzarsi su un'infinità di supporti differenti. È qui che nasce ufficialmente la tesi della realizzabilità multipla che, senza i concetti di imitazione e computazione per come sono stati sviluppati a partire da Turing, non avrebbe potuto realizzarsi almeno nella versione accolta dalle scienze cognitive classiche (cfr. Putnam 1960).

Ora, abbiamo già accennato che il contributo della macchina computazionale è stato fondamentale per l'emersione della robotica nei decenni successivi alla costituzione dell'impresa dell'intelligenza artificiale, ma lo stesso è stato detto per la macchina cibernetica. Quale concetto di macchina, quindi, è stato quello che ha informato e condizionato la nascita della robotica come disciplina?

### **1.3 Macchine: modelli per la mente**

Non sorprenderà sapere che, nonostante le ingenti differenze già sottolineate, i due concetti di macchina esposti non si escludono a vicenda e, anzi, costituiscono entrambi i precedenti necessari per la possibilità dell'emersione della robotica negli anni '70 nel XX secolo. Senza le riflessioni provenienti dal gruppo cibernetico, e le relative e numerose realizzazioni di macchina la cui utilità era legata a doppio filo all'opera di conoscenza, la robotica non avrebbe potuto acquisire lo status di disciplina a cavallo tra l'ingegneria e le scienze cognitive, al pari dell'intelligenza artificiale nel

momento in cui le tecnologie lo avrebbero reso possibile (Bernstein 1978/1982); senza la macchina computazionale, che fornisce la base su cui le teorie cognitive successive verranno sviluppate e provate, implementate e modificate, la robotica avrebbe perso il suo status di disciplina all'avanguardia, agli avamposti della ricerca scientifica sull'intelligenza artificiale e considerata tra le più sfidanti per la comprensione della mente umana. C'è da precisare, in ogni caso, che il panorama interdisciplinare in cui si configura il campo di studi dell'intelligenza artificiale influenzerà enormemente la costruzione dei primi robot: l'approccio simbolico-computazionale viene assunto quasi unanimemente a paradigma della nuova disciplina della scienza cognitiva in connubio con l'intelligenza artificiale, mentre la cibernetica si indebolisce fino quasi a ritirarsi sotto il peso delle critiche al primo connessionismo (ad es. Minsky & Papert 1959; cfr. Froese 2010; Damiano 2010).

La cifra distintiva della macchina computazionale viene apprezzata ancora meglio se messa in relazione, ancora una volta, alla macchina cibernetica rispetto al rapporto tra, da una parte, corpo, supporto, vita e, dall'altra, mente, cognizione, capacità. Come sottolineato da Boden (2006: 236):

The cyberneticists assumed that [...] not only is the mind a machine, but it essentially the same sort of machine as the living body is. The computationalists, by contrast, focused mainly on logical-linguistic thinking, problem solving, and purposeful planning. When they studied non-linguistic capacities, such as vision and locomotion, they assimilated them into logical thought. But they didn't assume that *metabolic* systems work by logic. On their view, then, the human mind is a machine that differs significantly from the biological machine in which it's embodied.

Il paragone mette in risalto non solo la grande differenza tra concezione del corpo nella cibernetica, legato alla vita e preso a oggetto di studio proprio nella sua duplice natura macchinica e organica, e nell'approccio simbolico computazionale, che vede il corpo quasi come un "ostacolo" alla computazione e identifica le sue caratteristiche nel pensiero logico. Infatti, ciò che risulta interessante è che:

- i) La macchina computazionale garantisce la sua applicabilità non più soltanto nella *equivalenza* di cervello e macchina, come pensato nella cibernetica, ma stabilisce un rapporto di *equiparazione* tra pensiero e algoritmo. Ciò che diventa centrale non è tanto l'operazione di costruire nuove macchine la cui struttura fisica potesse incarnare la funzione oggetto di indagine, quanto la capacità di programmare la macchina affinché raggiunga la capacità di portare a termine la computazione producendo il risultato esatto (Turing 1950; cfr. Boden 2006: 716).
- ii) Il metodo generato dall'introduzione della macchina computazionale viene declinato attraverso una trasformazione logica della maggior parte delle funzioni svolte da un agente intelligente. Anche quando sono le funzioni non propriamente logiche ad essere prese in considerazione, come la locomozione, la visione o la manipolazione di oggetti,

l'approccio simbolico-computazionale li traduce nel linguaggio logico, strategico e sintattico, scommettendo che sia questa la modalità principale attraverso cui il corpo porti avanti le funzioni elaborate dal pensiero.

- iii) Il rapporto metaforico tra macchina ed essere umano si trasforma in modo significativo. Se l'obiettivo della cibernetica era la meccanizzazione dell'umano, che si declinava nella comprensione dei meccanismi del pensiero e del comportamento attraverso la costruzione di macchine che incorporavano in un rapporto stretto il modello funzionale ipotizzato, la corrente simbolico-computazionale mette in moto il processo di antropomorfizzazione della macchina. La metafora diventa quindi biunivoca: il rapporto che fin da Cartesio aveva avuto una direzione che andava dall'umano alla macchina, stabilendo una connessione volta a inquadrare l'essere umano e il suo pensiero in ottica meccanicista, in questa sede torna su sé stesso, arrivando ad attribuire alla macchina caratteristiche umane, tra cui la capacità di un pensiero intelligente. La domanda posta da Turing (1950), se le macchine possano pensare, può considerarsi la prima attestazione di questa nuova dotazione di senso. Nella sua versione forte, se l'umano è una macchina, allora anche la macchina può essere umana.

Quello a cui siamo arrivati in questo momento della nostra ricostruzione, dunque, si configura come una macchina (ancora una volta) particolare: dalle radici storiche antichissime, ma allo stesso tempo nella sua originalissima configurazione, la macchina che si appresta a fornire il background tecnologico e il supporto materiale per la costituzione della robotica è un misto delle tradizioni analitiche, filosofiche e ingegneristiche incontratesi a cavallo della seconda metà del XX secolo. Il focus di indagine incentrato sul pensiero logico-linguistico, la risoluzione dei problemi e la pianificazione diretta a uno scopo, insieme ai progressi realizzati dai calcolatori, hanno instradato la via della robotica sotto la stella, allora brillante, dell'account simbolico-computazionale. Il retaggio cibernetico si manifesterà comunque nella robotica, come perseguimento della terza via del metodo sintetico (come nel caso di SHAKEY, *infra* §2.1): il ruolo che la costruzione dei robot acquista nella verifica e ottimizzazione delle teorie cognitive è innegabile, e anzi comporterà nel tempo anche qualche ripensamento (*infra* §2.2).

Fino a questo momento, abbiamo analizzato i presupposti che hanno condotto alla possibilità di pensare a una macchina capace di incarnare le caratteristiche peculiari dell'essere umano e del funzionamento della sua mente. Per comprendere il ruolo che i robot ricoprono o possono ricoprire oggi all'interno della sfida alla comprensione della mente umana, quello che ci spetta ora è un'indagine storica e filosofica del modo in cui l'ambizione dell'umano è stata declinata nella



costruzione di artefatti robotici allo scopo di spiegare il funzionamento dei processi cognitivi e come l'operazione di traduzione delle ipotesi circa il funzionamento della mente nelle architetture cognitive robotiche abbia comportato una rivoluzione delle scienze cognitive proprio a partire da riflessioni sull'efficacia dei comportamenti esibiti dai robot.

Arriveremo nel corso del prossimo capitolo a descrivere le caratteristiche del primo vero robot, SHAKEY, e a vedere come a partire dalla sua analisi sia stato possibile operare un cambiamento radicale sia all'interno degli studi sull'artificiale, sia in merito alle idee su mente e cognizione. Questa disamina non potrebbe essere realizzata se prima non si delineano i presupposti epistemologici delle teorie riguardo la mente e le funzioni cognitive provenienti da diverse discipline, che sono confluite nello sviluppo delle architetture robotiche e nella sfida rinnovata di comprendere l'essenza dell'essere umano, stavolta attraverso macchine cibernetico-computazionali. L'indagine, dunque, si costituirà come un viaggio di andata e ritorno: nel primo tragitto, partiremo dalle formulazioni di teorie riguardo al funzionamento della mente all'interno delle scienze cognitive e analizzeremo il modo in cui tali ipotesi sono state utilizzate per costruire artefatti robotici che funzionassero secondo gli stessi principi; la seconda parte prenderà le mosse dalla valutazione delle prestazioni realizzate dai primi robot, e si comporrà di un viaggio di ritorno dalla robotica alle teorie della mente e della cognizione nel campo delle scienze cognitive. Così, sarà possibile mostrare il ruolo della robotica del processo di definizione delle idee riguardo a cognizione, percezione e mente, puntando ad aprire nuovi orizzonti teorici e modalità di considerare il robot se messo in relazione con l'umano suo costruttore.

## Capitolo II – Dalla robotica alle scienze cognitive: andata e ritorno

Man is a creature who makes pictures of himself  
and then comes to resemble the  
picture.  
(Murdoch 1999: 75)

### 2.1 Viaggio di andata: dalle scienze cognitive alla robotica

#### 2.1.1 GOFAI: la ricetta per costruire robot

Negli anni successivi alla costituzione dell'Intelligenza artificiale (McCarthy et al. 1955/2006), il panorama che si delinea perseguendo l'obiettivo di indagare i meccanismi e le funzioni mentali sotto la guida dell'approccio simbolico-computazionalista è duplice (Boden 2014): da una parte, la prospettiva psicologica viene declinata in ottica computazionalista, traendo spunto dal funzionamento dei primi calcolatori per identificare le caratteristiche principali di funzioni psicologiche come memoria, percezione, attenzione, pensiero e linguaggio nel programma d'azione<sup>45</sup> (Boden 2006: 366-513); dall'altra, la componente ingegneristica, dedicata alla realizzazione di macchine intelligenti, era molto attiva e cominciava a produrre i suoi frutti con l'elaborazione di programmi per calcolatore molto avanzati, come il Logic Theorist (LT, Newell e Simon 1956) e il General Problem Solver (GPS, Newell et al. 1959), e quelli che sarebbero diventati i primi sistemi esperti (Collins 2012). Il legame tra le due spinte, quella psicologica e quella ingegneristica, era strettissimo e si manifestava con reciproche contaminazioni ed elaborazioni di ipotesi sia per quanto riguardo il funzionamento della mente, sia per la costruzione di macchine intelligenti. Ciò che teneva insieme le diverse realizzazioni del paradigma simbolico-computazionalista è la metafora della mente macchina, ma una macchina precisa: il computer.

La metafora della mente come computer è stata per decenni la cifra delle indagini svolte dalle scienze cognitive, dalla psicologia e, ovviamente, nell'ambito dell'intelligenza artificiale (Gigerenzer e Goldstein 1996; Boden 2006). Per tutte le versioni, la correlazione tra l'umano e il computer si articola sulla divisione tra hardware e software: la mente sta al software come il corpo sta all'hardware (Miller 1956; Fodor 1975; Searle 1990, 1991; Block 1993, 1995). In altre parole, la mente è vista come un elaboratore di informazioni, un computer in grado di ricevere input ed elaborarli, producendo output in uscita. Come già anticipato, questa formulazione viene identificata da Searle (1980) con una versione forte dell'Intelligenza Artificiale: la mente è un programma per computer e, di

---

<sup>45</sup> Margareth Boden ricostruisce le radici della psicologia cognitiva anche attraverso una disamina dell'apporto fondamentale che la psicologia nel *New Look*, elaborata tra gli altri da Jerome Bruner e Richard Gregory, ha fornito alla rivoluzione cognitiva della metà del secolo scorso (Boden 2006: 298-311).

conseguenza, qualsiasi sistema adeguatamente programmato è assimilabile alla mente umana, a prescindere dal supporto su cui è implementato.

In psicologia, il focus sul funzionamento del computer come chiave per comprendere i meccanismi della mente umana si configura come alternativa forte al comportamentismo, che in quel momento forniva un apparato teorico molto popolare per lo studio del comportamento umano e animale. Reagendo quindi all'idea comportamentista che le azioni e i comportamenti possano essere visti come catene di risposte condizionate da input esterni (Boden 2006: 257-262), uno dei primi contributi alternativi proviene dallo psicologo e fisiologo Karl Lashley. Lashley (1951) vede il comportamento intelligente come controllato da un programma centralizzato e gerarchicamente organizzato, e non più linearmente come un'unica risposta consecutiva alla ricezione di uno stimolo in entrata. Come sottolineato da Bermudez (2020: 22), i semi gettati da Lashley negli anni '50 germoglieranno in fretta nella psicologia computazionale, arrivando a influenzare profondamente l'indagine sulla mente e i tentativi di riprodurla. In particolare, due sono i principi fondamentali nella sua trattazione:

- i) L'idea che la maggior parte delle azioni intelligenti sia realizzata sotto il controllo di meccanismi di pianificazione e di processazione centralizzata dell'informazione, i quali operano però sotto la soglia di coscienza. Tale principio prende il nome di *Hypothesis of subconscious information processing*;
- ii) L'*Hypothesis of task analysis*, la quale formula una procedura secondo cui è possibile comprendere un *task* complesso scomponendolo in una gerarchia di *subtask* più basilari. Si ipotizza, quindi, che si possa arrivare a una gerarchia di *subtask* non più riducibile, che identifica quindi la struttura basilare del comportamento intelligente in oggetto.

Entrambi i principi entreranno a far parte dell'armamentario teorico di quella che Haugeland (1985: 112-113) definirà *Good Old Fashion Artificial Intelligence*, o GOFAI. Per il filosofo, nella GOFAI era implicito un riferimento alla versione forte dell'intelligenza artificiale, anche se non sempre gli esponenti di tale approccio hanno dichiaratamente espresso il loro favore (cfr. Boden 2006: 702). Si può constatare, infatti, che a unire sotto questa etichetta gli esponenti provenienti sia dal ramo psicologico, sia da quello ingegneristico, era una fiducia promettente nella metafora della mente come computer e nelle possibilità che questa concedeva nella costruzione di artefatti intelligenti. Tale prospettiva si manifestava soprattutto nella presa in carico di determinati oggetti di studio, tutti riassumibili sotto l'etichetta di *high-level human thought* (Boden 2014: 92). In particolare, i concetti centrali della GOFAI, come quello di ricerca euristica e pianificazione (Newell et al. 1958), venivano messi al centro dell'indagine ed erano pensati essere i meccanismi basilari di ogni processo cognitivo umano. Sotto l'etichetta 'processi cognitivi', tuttavia, non rientravano solamente task identificati

avere come caratteristica principale la logica sintattica, come il gioco degli scacchi, i meccanismi di decisione e di risoluzione di problemi matematici, ma anche compiti che a prima vista non sembravano coinvolgere attività logiche strutturate. Come sottolineato dal principio i), l'idea che la maggior parte dei meccanismi intelligenti sia svolta sotto il controllo di meccanismi di pianificazione o processazione centralizzata dell'informazione era valida per ogni compito portato avanti da un agente intelligente, anche se questi meccanismi non erano coscienti per l'agente. Secondo Boden (2014: 92),

This is partly because [GOFAI programs] can represent propositions with specific semantic content, and partly because they can include top-down executive control and self-monitoring processes. Even when they model unconscious processes, they typically represent these as being similar in kind to conscious deliberations. Visual recognition in GOFAI, for example, is often taken to be analogous to problem solving in logic. And logical problem solving, a fortiori, is seen as logic all the way down.

Questa idea porterà a sostanziali avanzamenti nell'intelligenza artificiale simbolica, con programmi come i dimostratori di problemi matematici e più avanti con i sistemi esperti (come MYCIN, Shortliffe 1977, e DENDRAL, Lindsay et al. 1980) e l'approccio della descrizione basata sulla descrizione processuale umana elaborato da Newell e Simon (1972). È chiaro che il secondo tipo di meccanismi presi in carico dall'analisi, come quello della visione suggerito da Boden, porta a una maggiore difficoltà di riproduzione in artefatti intelligenti, soprattutto perché implicano necessariamente la presenza di un apparato fisico che possa raccogliere dati input dall'ambiente, a differenza degli input per programmi al calcolatore che vengono consegnati al programma dai suoi sviluppatori.

L'esempio riportato da Boden riguardo la visione, infatti, è emblematico della questione che è necessario affrontare per poter costruire un robot: come è possibile rendere una macchina capace di elaborare i dati che provengono dall'esterno e farle produrre dei risultati corretti? La ricerca riguardo le operazioni logiche esplicite, come la costruzione di una strategia, della ricerca nello spazio del problema, della pianificazione, o della risoluzione di problemi tramite euristiche sembra portare a risultati strabilianti se implementata in programmi per computer. Ma sarà lo stesso per meccanismi diversi, come la visione e la manipolazione di oggetti nel mondo reale? La scommessa della GOFAI consiste proprio nel credere che anche task diversi da quelli esplicitamente logico-razionali possano essere scomposti con la stessa metodologia e realizzati in modo performativo su artefatti artificialmente intelligenti: la chiave era riuscire a scomporre attraverso il linguaggio formale della logica del primo ordine (First Order Predicate Logic) la struttura del task da portare a termine, che questo comprendesse una struttura logica di base o meno (Hertzberg 2007).

Nella filosofia della mente e nelle scienze cognitive degli anni '50 e '60, la domanda sulla possibilità di comprendere il modo attraverso cui l'informazione viene elaborata dall'agente

cognitivo prende le forme del problema mente-corpo, ovvero il problema di come spiegare il fatto che le esperienze, le percezioni, i pensieri o altri stati mentali si manifestino fisicamente nel cervello, considerato il luogo in cui tutta l'informazione raccolta convoglia e viene elaborata (cfr. Crane 2003; Howard 2023). Il problema, in realtà, affonda le sue radici agli albori della filosofia (Quintili 2024), ma ritorna in questo periodo con forza soprattutto nel contesto della costruzione di macchine intelligenti e con la ripresa del dualismo cartesiano (Crane 1999). I tentativi di risoluzione sono stati molteplici (cfr. von Bertalanffy 1964; Fodor 1981; Hanna e Thompson 2003) e hanno riguardato non solo il modo in cui è possibile per una macchina carpire le informazioni necessarie nell'ambiente in cui è immersa, ma pongono degli interrogativi anche ad una filosofia che deve fornire la base giustificatoria del processo di conoscenza. Ancora oggi, questo problema sembra non aver esaurito il suo potere inquisitorio (Lavazza 2008; Rowlands 2009; Nannini 2023).

Per rintracciare le radici della costruzione del primo robot, è importante soffermarci su una in particolare delle ipotesi elaborate per colmare il divario tra mente e corpo, dal momento che fornisce gli spunti per l'impianto teorico che verrà implementato nel primo robot ufficialmente mai realizzato.

#### *2.1.1.1 L'ipotesi del Sistema Fisico di Simboli*

Nel 1975, Allen Newell e Herbert Simon, nel loro celebre discorso alla consegna del Turing Award, sintetizzano la loro ipotesi del sistema fisico di simboli, elaborata per comprendere come l'informazione sia processata da un sistema e come sia possibile per questo raggiungere l'intelligenza. Il simbolo, nella loro formulazione, rappresenta una legge di strutture qualitative (*Laws of Qualitative Structures*) equiparabile al ruolo fondativo svolto dalla cellula in biologia, dalle placche tettoniche in geologia o la *germ theory of disease* elaborata da Pasteur: un elemento senza il quale la rivoluzione attestata nelle scienze dell'artificiale non avrebbe avuto luogo. Il simbolo assume dunque un'importanza centrale nella costruzione dell'intelligenza artificiale, dal momento che la capacità di registrare, conservare e manipolare simboli è considerata uno dei fondamentali «structural requirments for intelligence» (Newell e Simon 1976: 290). In particolare, il simbolo è identificato come un pattern fisico, come delle lettere scritte su un foglio, che possono essere combinati per formare strutture complesse, come le lettere possono essere combinate per formare delle parole e poi delle frasi, e così via. La struttura fisica di simboli può essere quindi manipolata e trasformata secondo regole. Nella definizione di Newell e Simon:

A physical symbol system consists of a set of entities, called symbols, which are physical patterns that occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structure. Besides these structures, the system also contains a collection of processes that

operate on expressions to produce other expressions: processes of creation, modification, reproduction and destruction. A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just symbolic expressions themselves. (Newell e Simon 1976: 292)

Prima di passare alle procedure di manipolazione di sistemi di simboli per creare nuove espressioni nei sistemi artificiali, ovvero del procedimento delle ricerche euristiche, ciò che è importante sottolineare è che il sistema fisico di simboli viene assunto come base per una ipotesi sulla natura e sul funzionamento dell'intelligenza. Infatti, come affermano gli stessi autori, «A physical symbol system has the necessary and sufficient means for general intelligent action» (Ibid.: 293). Con “necessario”, gli autori intendono dire che ogni sistema che esibisce intelligenza può essere provato essere un sistema fisico di simboli; con “sufficiente”, intendono affermare che qualsiasi sistema fisico di simboli «of sufficient size» (Ivi) può essere organizzato in modo tale da produrre azioni intelligenti, ovvero azioni che se svolte da un essere umano non esiteremmo a definire intelligenti. Ciò viene reso possibile nel sistema fisico di simboli perché esso contiene anche i processi per manipolarli: in questa formulazione, pensare non è altro che trasformare le strutture di simboli in altre strutture di simboli in accordo a determinate regole (Bermudez 2020: 103). Lo sfondo teorico che sorregge tali affermazioni consiste nell'idea, anche questa elaborata nel dettaglio da Newell e Simon (1972), che il pensiero intelligente si configuri come l'abilità di risolvere problemi: se, quindi, scomponiamo il procedimento di risoluzione dei problemi tramite linguaggio formale, saremo in grado di riprodurre tale funzionamento utilizzando un linguaggio di programmazione che garantisce possibilità di *denotazione e interpretazione* nei programmi per computer<sup>46</sup>.

L'ultima caratteristica da sottolineare che è posseduta dal sistema fisico di simboli elaborato da Newell e Simon è infatti proprio la capacità del sistema stesso di rappresentare contenuti proposizionali. Il sistema ingloba il funzionamento della logica proposizionale, inventato da Boole e già impiegato nel funzionamento dei programmi per calcolatore da Claude Shannon e John McCarthy (Davis 2000). Tale sistema consiste nell'elaborazione di un metodo formale per manipolare e analizzare proposizioni logiche utilizzando operatori come AND ( $\wedge$ ), OR ( $\vee$ ), NOT ( $\neg$ ), implicazione ( $\rightarrow$ ), bicondizionale ( $\leftrightarrow$ ), e così via. Già alcune realizzazioni della macchina di Turing avevano la possibilità di incorporare metodi per la manipolazione di funzioni della logica proposizionale, e nel sistema fisico di simboli questo si concretizza come la possibilità di stabilire regole di interpretazione

---

<sup>46</sup> Nell'impresa della costruzione di macchine intelligenti, il sistema fisico di simboli viene istanziato attraverso la creazione di un linguaggio che consente al sistema di simboli la capacità di *denotazione*, ovvero la garanzia di accesso all'oggetto per mezzo dell'espressione presa in carico, e di *interpretazione*, cioè la capacità del sistema di realizzare il processo denotato dall'espressione presa in carico (Newell e Simon 1976: 292). Nella fattispecie, gli autori riportano l'esempio di LISP (LISt Processing language), un linguaggio di programmazione elaborato da McCarthy nel 1958 considerato il primo linguaggio di programmazione funzionale, il linguaggio di programmazione più utilizzato per i trenta anni successivi (Russell e Norvig 2010: 19).

che a loro volta consentono di assegnare un significato (e quindi un contenuto) alle espressioni coinvolte.

Il concetto di rappresentazione diventa cruciale, allora, non solo per rendere capace la macchina di interpretare il sistema fisico di simboli e le espressioni, ma anche come soluzione adottata rispetto al problema mente-corpo, o mente-mondo. Infatti, quello di rappresentazione è un concetto sviluppato per garantire alla mente di avere accesso al mondo (di cui il corpo fa parte). Il concetto è fondamentale per le scienze cognitive, dai suoi albori fino ad arrivare al giorno d'oggi (Shea 2018), e risulta troppo ampio per essere trattato in questa sede. Per la nostra argomentazione, in ogni caso, è sufficiente ricorrere a una definizione standard, utile a illustrare il modo in cui il sistema fisico di simboli costituisce una rappresentazione del mondo reale. Infatti, l'ipotesi avanzata da Newell e Simon, che riassume quella della GOF AI in generale, si basa sull'idea che per svolgere un qualsivoglia compito cognitivo, dal ragionamento al movimento, dal problem-solving alla visione alla locomozione, sia necessario per l'agente *rappresentarsi* il mondo in qualche modo. *Rappresentare* significa che le entità del *mondo reale*, insieme alle loro relazioni e proprietà, possono essere modellate e trattate attraverso l'uso di simboli, ovvero rappresentazioni. L'attività intelligente è quindi realizzata sia negli agenti biologici sia in quelli artificiali come la capacità di 'codificare' la conoscenza riguardo al mondo esterno (compreso il corpo, o hardware) attraverso rappresentazioni simboliche astratte interne alla mente (Lieto 2021: 10). La cognizione non consiste in nient'altro se non nell'attività di manipolazione *fisica* di queste rappresentazioni mentali (Shea 2018: 4), assimilata quindi al concetto di computazione (Pylyshin 1980).

In sostanza, quello che fa di un agente un agente intelligente è la sua capacità di manipolare simboli e di portare a termine il compito di risoluzione del problema attraverso l'impiego di regole specifiche accettate dal sistema stesso. Una particolare caratteristica di questi sistemi è relativa ai metodi di manipolazione delle espressioni prese in carico, fondamentale nel procedimento di risoluzione dei problemi: la ricerca euristica.

### 2.1.1.2 Ricerca euristica

La ricerca empirica di Allen Newell e Herbert Simon riguardo gli aspetti formali dell'intelligenza come il problem-solving si concretizza nella costruzione di programmi per la risoluzione di problemi come i già ricordati Logic Theorist e General Problem Solver. I meccanismi che guidavano il tentativo di risoluzione dei problemi in questi programmi erano un esempio di quello che gli autori definivano tecniche di ricerca euristica (*Heuristic Search Techniques*), ovvero strategie per la ricerca delle modalità di risoluzione dei problemi. È attraverso queste tecniche che i sistemi fisici di simboli

riescono a portare avanti processi di risoluzione di problemi. Nella loro definizione, la *Heuristic Search Hypothesis* viene così formulata:

The solutions to problems are represented as symbol structures. A physical symbol system exercises its intelligence in problem solving by search – that is, by generating and progressively modifying symbol structures until it produces a solution structure. (Newell e Simon 1976: 300)

I concetti fondamentali per comprendere l'utilità della ricerca euristica nella risoluzione dei problemi sono quello di *spazio della ricerca* (search space) e quello di *stato*. Lo spazio della ricerca (nella risoluzione del problema) è identificabile con l'insieme delle possibilità da intraprendere da parte del sistema per risolvere il problema: in una partita di scacchi, ad esempio, in un dato momento della partita, il giocatore avrà un numero finito di mosse da poter giocare per avvicinarsi al suo obiettivo, ovvero lo scacco matto al re avversario; d'altra parte, lo stato consiste nella determinata situazione in cui il sistema si trova in un dato momento del tempo nel processo di risoluzione: da uno stato iniziale, il sistema può applicare un insieme di regole di trasformazione dello stato per avvicinarsi allo stato identificato come finale. Lo spazio della ricerca consiste quindi di tutti gli stati che possono essere raggiunti dallo stato iniziale applicando le trasformazioni permesse.

Il modo più diffuso per rappresentare lo spazio della ricerca da parte degli informatici e dei ricercatori in questi processi è la struttura ad albero, che rende più semplice la visualizzazione del problema e che permette di visualizzare contestualmente le diverse opzioni percorribili in un determinato stato del processo. Si può dire che il sistema abbia risolto il problema quando ha trovato lo stato di soluzione nello spazio della ricerca. Per raggiungere questo stato, diverse euristiche entrano in gioco e possono essere applicate a seconda del task preso in esame. Nel caso del GPS, ad esempio, Newell e Simon avevano elaborato una serie di tecniche di ricerca euristica che prendono il nome di *means-end analysis*. In particolare, la *means-end analysis* consisteva in quattro semplici istruzioni per il programma (Bermudez 2020: 105):

1. Valutare la differenza tra lo stato corrente in cui l'agente si trova e lo stato della soluzione da raggiungere.
2. Identificare la trasformazione che riduce la differenza tra lo stato corrente e lo stato della soluzione.
3. Controllare che la trasformazione in (2) possa essere applicata allo stato corrente.
  - a. Se sì, applicare la trasformazione e tornare al punto (1).
  - b. Se no, ritornare al punto (2).

Il processo di ricerca della soluzione, dunque, consiste nel valutare la situazione in cui il sistema si trova nel momento corrente, metterlo in relazione all'obiettivo desiderato, e calcolare la mossa



migliore per avvicinare lo stato corrente allo stato della soluzione. La ricerca euristica consente di ridurre lo spazio della ricerca affinché il sistema possa esplorare dei percorsi senza incappare nell'esplosione combinatoria (spesso causate da *brute force searches*, ovvero ricerche che esplorano ogni ramo aperto dallo spazio della ricerca, che possono funzionare solamente per problemi molto semplici), gerarchizzando i livelli e consentendo una esplorazione ridotta, o sub-ottimale. Quando lo stato corrente e lo stato di soluzione combaceranno, si potrà dire che la soluzione è stata trovata.

In sostanza, i più saldi punti di forza dell'approccio elaborato da Newell e Simon consistono nel rendere il sistema capace di modellizzare la gerarchia e l'ordine sequenziale, di permettere la precisione nel problem-solving e di rappresentare specifici contenuti proposizionali. Ciascuna di queste caratteristiche è mostrata nel programma SHRDLU (Winograd 1972).

### 2.1.1.3 Logica del razionale: SHRDLU

Prendiamo quindi come esempio il programma SHRDLU, elaborato da Terry Winograd per la sua ricerca di dottorato tra il 1968 e il 1970 al MIT, lo stesso posto in cui venivano sviluppati lavori pionieristici in robotica come quello della mano portato avanti da Fogel, McCulloch e Ramsey-Klee (1970: 256-265). SHRDLU era il primo programma di *Natural Language Processing* (NLP) capace di analizzare frasi in lingua inglese di un grado significativo di complessità, come ad esempio «How many eggs would you have been going to use in the cake if you hadn't learned your mother's recipe was wrong?» (Boden 2006: 686). Niente a che vedere con i programmi di NLP sviluppati fino a quel momento. Uno dei più celebri esempi di questi programmi scritti prima di SHRDLU è ELIZA: elaborato da Joseph Weizenbaum nel 1966, era capace di simulare una conversazione riformulando le frasi inserite dal dialogante umano nella fase precedente in forma di domande, imitando la struttura delle conversazioni tipiche della psicanalisi rogeriana<sup>47</sup>. Nonostante anche ELIZA sia stato uno dei programmi più di successo del periodo della GOFAI, le sue prestazioni erano molto più limitate e non andavano oltre la mera associazione di simboli linguistici per rispondere a input della stessa natura. Al contrario, il caso di SHRDLU è di notevole interesse in questa sede perché, nonostante il programma non sia un robot fisico reale, le attività che poteva realizzare comprendevano, oltre che il NLP, anche la simulazione di un braccio robotico che aveva lo scopo di individuare e spostare dei blocchi geometrici simulati nel suo micromondo<sup>48</sup>. Infatti, lo scopo per cui il programma è stato

---

<sup>47</sup> Esempi di conversazione con ELIZA possono essere trovati in Weizenbaum (1966: 36-37) e Boden (1977: 106-107). Si può anche provare a conversare direttamente con ELIZA, in una sua versione aggiornata, al sito: <https://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm>

<sup>48</sup> Il task svolto da SHRDLU e le sue attività simulate in un mondo virtuale hanno diversi antenati illustri. Tra i tanti, è interessante menzionare i tentativi di costruzione di un braccio robotico collegato a una videocamera e a un computer di Marvin Minsky, all'epoca al MIT, che cercava di dare occhi e mani a delle menti artificiali così che le macchine avessero la capacità di «see, plan, and act» (Minsky 1985: 44-45).

realizzato era, sì, quello di manipolare il linguaggio naturale, ma per fargli *comprendere* delle istruzioni in lingua inglese riguardo azioni che doveva intraprendere con i blocchi, e che lo facesse in tempo reale (Winograd 1972: 7-8).

Per rendere SHRDLU capace di capire le istruzioni che gli venivano impartite, quindi, doveva essere elaborato un programma per l'analisi del linguaggio naturale. A tal proposito, Winograd è stato uno dei primi a suggerire l'impiego di una semantica procedurale per costruire programmi capaci di analizzare il linguaggio umano. In dettaglio, la semantica procedurale cerca di rappresentare il significato linguistico secondo un'ipotesi semantica di come la mente (o il sistema di intelligenza artificiale) effettivamente lo processa: le parole e le frasi sono considerati come dei mini-programmi, un insieme di istruzioni date all'ascoltatore o ricevente del messaggio da impiegare per cercare o stabilire determinate rappresentazioni nella mente e performare quindi certe operazioni computazionali o inferenziali su queste rappresentazioni (Boden 2006: 412). Nonostante Winograd avesse esplicitamente dichiarato che SHRDLU non voleva essere una rappresentazione realistica dei processi psicologici e linguistici dell'essere umano, le performance realizzate dal programma lo portarono all'attenzione di psicologi e filosofi, stupiti dalle abilità di 'comprensione' esibite<sup>49</sup>.

Ma cosa riusciva a capire SHRDLU? Le domande che il programma sapeva analizzare e che gli fornivano effettivamente delle istruzioni riguardavano le azioni che il programma poteva intraprendere manipolando un certo numero di diversi blocchi. Nella figura seguente (Fig. 1.1) è possibile osservare una configurazione standard all'interno di SHRDLU.

---

<sup>49</sup> Tanta fu l'attenzione nei confronti di SHRDLU, che gli fu dedicato un intero numero della rivista *Cognitive Psychology*. Nell'introduzione al volume, i curatori tengono a precisare: «Winograd's system is not a "simulation", but it incorporates important ideas about human syntactic, semantic, and problem-solving abilities, and, in particular, about their interactions in understanding natural language. Human intelligence goes far beyond this system, and human language comprehension may turn out to differ in major respects from the means employed here. At the very least, however, Winograd's system should prove an invaluable tool for thinking about what we do when we understand and respond to natural language» (Winograd 1973: vii).

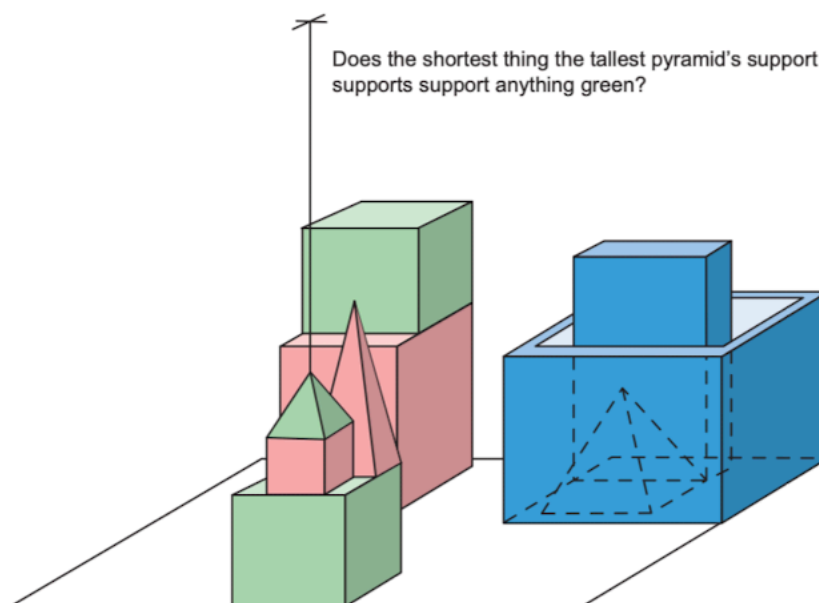


**Figura 1.1** Una domanda per SHRDLU riguardo il suo micromondo virtuale. (Bermudez 2020: 40)

Come è possibile notare, le domande rivolte al programma sono relative alla configurazione dei blocchi nel micromondo virtuale su cui SHRDLU può agire. Il concetto di micromondo<sup>50</sup>, opposto ai programmi di linguaggio *general-purpose* come ELIZA, risulta di fondamentale importanza per il funzionamento del programma e, in generale, si rende necessario affinché sia possibile individuare i blocchi e capire il modo in cui possano essere manipolati. Infatti, il mondo di SHRDLU è un ambiente simulato in cui il sistema può operare attraverso una serie di regole definite, dimostrando quindi anche capacità di interazione con gli oggetti virtuali che lo popolano.

Il micromondo di SHRDLU è molto semplice: ci sono una serie di blocchi colorati che riproducono diverse figure geometriche, come cubi, piramidi e parallelepipedi, le quali sono posizionate su un piano. Il braccio robotico nella facoltà di SHRDLU era in grado di afferrare i blocchi, spostarli, muoverli, e se istruito metterli in una scatola. Se visto oggi, SHRDLU non sembra molto sofisticato, ma all'epoca consisteva nella prima realizzazione di un programma capace di interagire in tempo reale con gli oggetti del suo mondo virtuale. SHRDLU ha costituito uno degli esempi più riusciti della GOF AI e per le scienze cognitive ha segnato un momento fondamentale, almeno per tre ragioni (Bermudez 2020):

- i) Fornisce un esempio di come regole e principi astratti come quelli incorporati nella semantica e nella sintassi procedurale potessero essere implementati empiricamente in programmi per calcolatore. Nonostante il dichiarato rifiuto della componente sintetica,

<sup>50</sup> Il termine “micromondo” tornerà di nuovo in questa trattazione, ma con diversa accezione, in §5 (cfr. Varela 1991b).

SHRDLU fornisce indirettamente una prova empirica della plausibilità delle ipotesi sviluppate riguardo alla comprensione e alla simulazione della comprensione del linguaggio naturale;

- ii) L'assunto fondamentale incorporato da SHRDLU consiste nell'ipotesi che la comprensione del linguaggio naturale sia descrivibile tramite un algoritmo. Come dichiarato da Winograd, «[a]ll language use can be thought of as a way of activating procedures within the hearer» (1972: 104). Il programma converte tramite rappresentazione simbolica gli input dati dall'istruzione e l'algoritmo in forma di euristica gli rende possibile compiere l'azione impartita. Nella struttura, non è niente di diverso rispetto a quello che faceva la macchina universale di Turing, ma nello specifico SHRDLU lo mostrava per la comprensione del linguaggio.
- iii) Si configura come la realizzazione del principio di scomposizione in sub-task per la risoluzione dei problemi, come adottato anche dall'approccio della ricerca euristica esposto poco sopra. In particolare, SHRDLU portava avanti distintamente e parallelamente il funzionamento di un sistema sintattico (*syntactic system*), uno semantico (*semantic system*), e uno cognitivo-deduttivo (*cognitive-deductive system*), le cui analisi venivano poi integrate alle informazioni riguardo al mondo già in suo possesso. Ciò era reso possibile perché tutte e tre le forme di conoscenza venivano rappresentate in modo simile, ovvero in termini di procedure (Bermudez 2020: 42). Il punto cruciale consisteva nel fatto che tutti i gradi di analisi e modellazione (sintassi, semantica, *reasoning*, e *world knowledge*) potevano interagire per l'analisi di una singola frase, portando avanti anche un processo di problem-solving per effettuare il giusto movimento dei blocchi seguendo l'istruzione (Boden 2006: 686). In questo caso, si dimostra anche l'importanza della gerarchia e dell'ordine sequenziale (pena il fallimento dell'istruzione).

I risultati raggiunti da SHRDLU erano per il tempo un grande passo in avanti anche per l'intelligenza artificiale: per la prima volta, venivano integrate attività come la comprensione a livello di frase, il ragionamento su contenuti di domini specifici, l'esecuzione di attività con più sotto-obiettivi e l'interazione naturale con l'utente umano. Le prestazioni erano raggiunte, tuttavia, grazie ad un alto grado di semplificazione relativo sia al micromondo di SHRDLU, sia alle regole di manipolazione. Infatti, il programma operava all'interno di un dominio ben strutturato e definito, le cui regole di manipolazione erano dal principio accessibili, e aveva piena conoscenza dello stato dell'ambiente simulato. E, tuttavia, non di rado il programma incappava in dei *bug*, rendendogli impossibile continuare il suo processo. In altre parole, non c'erano imprevisti, tutte le informazioni

che il programma doveva conoscere erano note e, fino a che veniva impartita la giusta istruzione, il programma era sempre in grado di portare a termine il proprio task. D'altronde, l'ambiente simulato e le strette regole pre-programmate garantiscono proprio questo, ma, come avrebbe in seguito sottolineato lo stesso Winograd, SHRDLU non aveva alcuna comprensione delle richieste che gli venivano poste<sup>51</sup> e nessuna capacità di improvvisare le modalità di risoluzione dei problemi. Al suo programma mancavano anche le più basilari capacità pragmatiche di interpretazione: fu questo uno dei motivi per cui Winograd cambiò radicalmente idea negli anni successivi, arrivando a criticare aspramente l'approccio della GOFAI che aveva contribuito a creare (Winograd e Flores 1986; *infra* §2.3.1.1).

Ora, abbiamo visto come l'approccio della GOFAI sia stato implementato su uno dei primi bracci robotici virtuali, costruito nei primi anni '70. Nello stesso periodo, alla Stanford University, un diverso gruppo di ricercatori stava portando avanti un progetto molto simile a SHRDLU negli intenti, ma questa volta con un robot in 'carne ed ossa'. SHAKEY, questo il suo nome, è considerato ufficialmente il primo vero robot autonomo programmabile mai costruito, e implementa in linea generale gli stessi principi che consentivano a SHRDLU di compiere le sue attività. Avrà, quindi, SHAKEY le stesse capacità di manipolare gli oggetti in tempo reale che aveva dimostrato SHRDLU?

### 2.1.2 Archeologia di SHAKEY

Abbiamo precedentemente affermato che la costruzione dei primi sistemi robotici non sia nata come un'impresa ingegneristica sulla base del metodo sintetico. Non si trattava, cioè, di costruire in modo esplicito un robot per convalidare le ipotesi avanzate in una teoria riguardo al funzionamento di un comportamento o di una funziona mentale, quanto invece di sfruttare le ipotesi allora considerate valide per costruire macchine che funzionassero. Anche se l'intento non era esplicito, dunque, sono state proprio le teorie del sistema fisico di simboli e del ragionamento per euristiche a fornire il supporto teorico per la costruzione del primo robot autonomo in grado di realizzare alcune azioni nel mondo.

SHAKEY nasce come progetto condotto allo Stanford Research Institute (SRI, oggi SRI International, cfr. Kuipers et al. 2017) da personalità come Charles Rosen, Nils Nilsson, e Peter Hart, con collaboratori come Bertram Raphael e Helen Chan Wolf, finanziato dalla Defense Advanced Research Projects Agency (DARPA, Nilsson 1984). Lo scopo dichiarato di questa impresa era «to develop concepts and techniques in artificial intelligence enabling an automaton to function

---

<sup>51</sup> «The AI programs of the late sixties and early seventies are much too literal [...]. This gives them a “brittle” character, able to deal well with tightly specified areas of meaning in an artificially formal conversation. They are correspondingly weak in dealing with natural utterances, full of bits and fragments, continual (unnoticed) metaphor, and reference to much less easily formalizable areas of knowledge» (Winograd 1976: 17).

independently in realistic environments» (Ibid.: 4). Tra il 1966 e il 1972, il gruppo di ricercatori sviluppa due versioni del sistema robotico: la prima, finalizzata nel 1969, consisteva in un veicolo mobile equipaggiato con una telecamera e sensori radiocomandati; la seconda, a cui arrivarono a compimento nel 1971, era un sistema robotico più potente del precedente, reso tale da miglioramenti significativi e sostanziali del programma incorporato e rimpiazzando i radiocomandi con versioni digitali di comunicazione (Kuipers et al. 2017). Il suo curioso nome deriva dal fatto che ciò che catturava l'attenzione quando in funzione era il suo andamento traballante e il tremolio nelle mosse: per questo, Charles Rosen suggerì il nome SHAKEY<sup>52</sup>.

Differentemente da SHRDLU, quindi, SHAKEY era un agente fisico, con una presenza reale nel mondo, da alcuni annoverato come primo esempio di *cognitive robotics* (Carlucci Aiello et al. 2012). Si può dire, infatti, che la costruzione di SHAKEY è il primo vero tentativo di costruire un agente “completo” adottando delle ipotesi sul funzionamento della mente, della cognizione e dell'intelligenza per come allora erano state formulate nel campo congiunto delle scienze cognitive e dell'Intelligenza Artificiale a trazione GOFAL. Proprio a causa di questo passaggio da mondo virtuale a mondo fisico, vedremo che i problemi che erano già emersi in SHRDLU (come l'inesistenza di una pragmatica, sia nel linguaggio che nell'azione, o l'incapacità di improvvisare o di fare riferimento al senso comune) si mostrano ancora più evidenti in SHAKEY, conducendo a riflessioni riguardo al ruolo e alle modalità della percezione, al rapporto tra percezione e azione, e alla configurazione generale della teoria dell'intelligenza e della mente per come era stata concepita nella GOFAL.

Vediamo, allora, nel dettaglio cosa era capace di fare SHAKEY e quali riflessioni ha suscitato all'interno dei campi disciplinari che lo hanno realizzato.

### 2.1.2.1 Il mondo di SHAKEY

In sostanza, l'hardware di SHAKEY consisteva in una struttura metallica, una serie di sensori che gli permettevano di estrapolare informazioni dall'ambiente circostante (sonar range finders, collision detector sensors, “cat-whisker” touch sensors, Nilsson 1984: 9), era posizionato su ruote alla base che gli consentivano di muoversi in uno spazio e aveva un'antenna all'estremità superiore che permetteva un collegamento radio o digitale. Il collegamento era fondamentale per SHAKEY: infatti, il software che permetteva alla struttura di muoversi non era posizionato localmente su SHAKEY, ma lavorava su un computer da lui separato, il quale inviava le istruzioni e i comandi comunicando attraverso

---

<sup>52</sup> «We worked for a month trying to find a good name for it, ranging from Greek names to whatnot, and then one of us said, ‘Hey, it shakes like hell and moves around, let’s just call it Shakey.’» (v. Shakey in *Exhibition: Artificial Intelligence & Robotics*. Computer History Museum. <https://www.computerhistory.org/revolution/artificial-intelligence-robotics/13/289>)

l'antenna. In sostanza, seguendo la metafora computazionale, possiamo dire che il cervello di SHAKEY si trovasse non nel suo corpo, ma in un computer nella stanza adiacente.

Come sottolineato in più sedi, SHAKEY è il primo robot mai costruito capace di muoversi in un ambiente, “percepire” informazioni da questo, seguire istruzioni e implementarne di complesse in un ambiente realistico. Il passo in avanti rispetto a tecnologie e programmi come SHRDLU è significativo: da un mondo virtuale in cui il braccio era a disposizione del programma ma non realizzato fisicamente, si passa a una versione in cui è possibile per la macchina manipolare attivamente lo scenario in cui si trova a performare. Il mondo di SHAKEY, tuttavia, è ancora un mondo realistico, non reale: il suo mondo era formato da un ambiente controllato, composto principalmente da una o una serie di stanze collegate tramite porte e da alcuni oggetti con cui interagire o ostacoli da evitare: in sostanza, un mondo giocattolo (o *toy world*<sup>53</sup>, Brooks 1986).

L'abilità principale data a SHAKEY dal suo programma era quello di lavorare su un ‘modello’ del mondo in cui il robot doveva agire (Nilsson 1984: 6). Nella figura che segue (fig. 1.2) è possibile vedere una rappresentazione del modello di mondo su cui opera il software di SHAKEY.

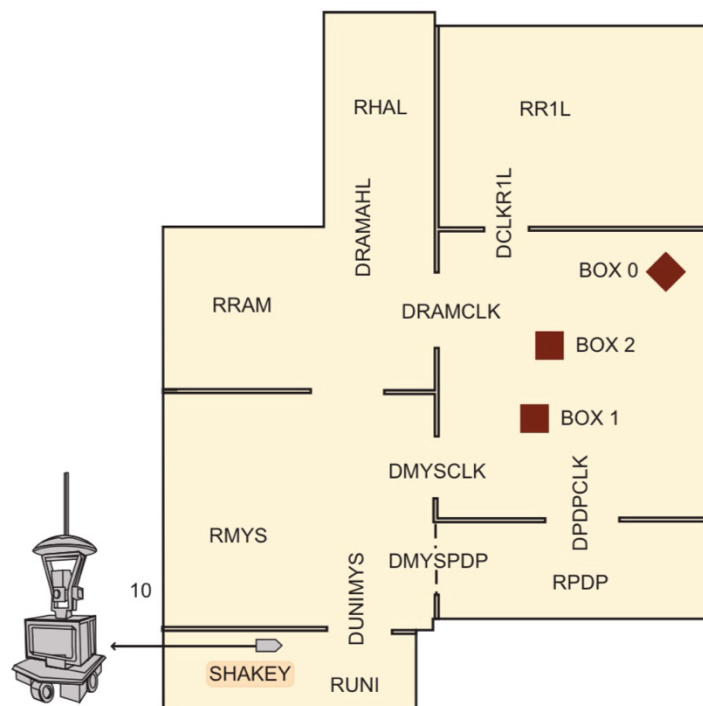


Fig. 1.2 Una mappa dell'ambiente fisico di SHAKEY (Bermudez 2020: 409, adattato da Nilsson 1984)

Ogni ambiente rappresentato nel mondo di SHAKEY ha un nome: le stanze sono tutte caratterizzate da nomi con l'iniziale R (come ad esempio RMYS, ovvero una stanza misteriosa, ‘mystery room’, di cui SHAKEY non ha alcuna informazione circa il suo contenuto, o RCLK, che

<sup>53</sup> «The definition of “toy” [world] includes all worlds where it is not up to the AI system itself to do all the understanding of the world itself without relying on a human interpreter» (Brooks 1986: 5).

identifica la stanza con l'orologio), le porte e i passaggi tutti con la D, specificando anche quali ambienti possono collegare (ad esempio, DUNIMYS identifica un passaggio tra la stanza in cui SHAKEY si trova nella fase di avvio del programma, RUNI, e la stanza RMYS). In questo mondo, SHAKEY poteva spostarsi da una stanza all'altra, ma anche raccogliere e spostare alcuni oggetti, evitare alcuni ostacoli (ad esempio blocchi posizionati lungo il suo percorso che avrebbero impedito al robot il raggiungimento dell'obiettivo), passare sopra ad alcune scatole (*box*), accendere o spegnere la luce, e attività di complessità simile (Nielssen 1984; Russel & Norvig 2010: 398).

Senza dubbio, la prospettiva interessante per l'analisi di SHAKEY non è tanto quella dell'hardware, che comunque risulta di notevole valore se visto nell'ottica di un avanzamento tecnologico materiale per l'implementazione dell'intelligenza artificiale, ma quella del software. Come sottolineato dagli stessi ricercatori (Kuipers et al. 2017), le idee più innovative sviluppate in questo robot riguardavano:

- i) le modalità di costruzione del mondo in SHAKEY, il quale veniva rappresentato da predicati «in the first-order predicate calculus» (Kuipers et al. 2017: 93). Grazie alla logica del primo ordine<sup>54</sup>, il mondo di SHAKEY rappresentava degli oggetti (come porte, blocchi, scatole e stanze) e le relazioni che intercorrono tra questi attraverso delle proposizioni in un vocabolario basilare e concedeva lo sviluppo delle operazioni logiche utili al suo funzionamento. Il vocabolario conteneva anche un nome per SHAKEY e dei predicati che descrivessero lo stato del robot, come la posizione nel suo mondo o la distanza dall'obiettivo (Bermudez 2020: 408).
- ii) la struttura del software di controllo di SHAKEY come una serie di strati, o *layer*, impiegati per la prima volta in un robot. Questa idea si rifà direttamente alla teoria secondo cui i comportamenti complessi siano il risultato di una pianificazione e una organizzazione precedenti, introdotta per la prima volta da Lashley. L'organizzazione gerarchica si configura, quindi, come abbiamo visto in precedenza, attraverso la scomposizione del piano generale in piani più semplici, fino ad arrivare alla struttura essenziale non più scomponibile a sua volta in sotto-piani. Ad esempio, per far sì che SHAKEY arrivi alla BOX-0, si procederà scomponendo il piano generale 'trova BOX-0' nei piani più semplici come il 'walking plan' e il 'research plan'.

Entrare nel dettaglio della suddivisione degli strati dell'architettura di SHAKEY ci permette di vedere come questa sia a tutti gli effetti una realizzazione robotica delle idee prima esposte riguardo

---

<sup>54</sup> Come accennato in precedenza, la logica o teoria del primo ordine, o calcolo dei predicati, è il linguaggio logico che permette di parlare a riguardo di determinati oggetti come aventi particolari proprietà, permettendo anche di formulare generalizzazioni circa gli oggetti e di denotare relazioni tra questi. Cfr Davis (2000)



il sistema fisico di simboli e il ragionamento euristico. Innanzitutto, la macro-divisione di questi strati, in tutto cinque, può essere raccolta in due gruppi. Il primo gruppo contiene solamente l'hardware level (Bermudez 2020: 409), ma è quello che fornisce l'ancoramento materiale dell'architettura del software al mondo reale, consentendo al robot di navigare e interagire fisicamente con un ambiente.

Il secondo gruppo è quello del software, che consta di quattro strati gerarchizzati di controllo. Il primo di questi, detto *Low-Level Action* (LLAs, Nielssen 1984: 25-33; Kuipers et al. 2017: 93), è lo strato fondativo del software che “parla” direttamente all'hardware di SHAKEY e fa riferimento alle sue capacità fisiche di interazione con l'ambiente e di movimento. In questo strato sono presenti i comandi per i comportamenti standard di SHAKEY, che possono essere considerati i blocchi irriducibili e non più scomponibili della gerarchia di comportamento. I comandi riguardano capacità di SHAKEY come il movimento orizzontale (PAN) o quello verticale (TILT) della testa, il movimento laterale per mezzo delle ruote incorporate (ROLL), scattare una foto tramite la camera (SHOOT) o di direzionarsi verso una specifica direzione (TURN) (Nielssen 1984: 28). Salendo nella gerarchia, il secondo livello del software è detto *Intermediate-Level Action* (ILAs), e si costituisce di una serie di condizioni e di azioni da svolgere una volta che le condizioni sono soddisfatte (Ibid.: 35-50). In sostanza, questo livello è il primo livello di organizzazione del comportamento di SHAKEY: tale livello di organizzazione consta sia di routine di azioni e operazioni, ovvero sequenze di LLAs che vengono richiamate per eseguire istruzioni date dal piano generale, come ‘raggiungere un'altra stanza’, ma ha anche la capacità di richiamare ulteriori ILAs (Ibid.: 42). Ad esempio, la action routine GETTO conduce SHAKEY in una stanza e, una volta arrivato, la action routine NAVTO richiamata permettere a SHAKEY di esplorare la stanza in cui si trova (Bermudez 2020: 411). Questo permette una velocità maggiore di pianificazione, in quanto la ‘discesa’ nella gerarchia è limitata alla convocazione di action routine che al loro interno gerarchizzano di già le LLAs necessarie al suo svolgimento. Tuttavia, come sottolineato da Kuipers e colleghi (2017: 93), se l'architettura di SHAKEY si fosse fermata qui non si sarebbero raggiunti la maggior parte dei traguardi che avrebbero portato in futuro a considerarlo una pietra miliare dello sviluppo dell'intelligenza artificiale. Infatti, l'architettura descritta fino a qui avrebbe permesso al robot di raggiungere obiettivi semplici, che prevedessero solamente una sequenza di azioni pre-programmata (Nielssen 1984: 65). Quello di cui si voleva rendere SHAKEY capace, però, era una pianificazione in tempo reale del processo di svolgimento della funzione e risoluzione del problema, la cui possibilità di implementazione viene rintracciata nella capacità di comporre sequenze di azioni per organizzarle in un piano generale a partire dalle informazioni raccolte nell'ambiente e dalla conoscenza in suo possesso riguardo all'obiettivo da raggiungere. Tale possibilità viene sviluppata con STRIPS.

#### 2.1.2.2 Livelli gerarchici di organizzazione: STRIPS

STRIPS (Stanford Research Institute Problem Solver) è un sistema, o linguaggio, di pianificazione che lega e gerarchizza diverse ILAs per raggiungere un determinato obiettivo (Nielssen 1984: 57). In sostanza, se SHAKEY doveva raggiungere una stanza passando attraverso porte, evitando ostacoli, con l'obiettivo di muovere una scatola, STRIPS gli concedeva la capacità scomporre il problema legato al raggiungimento dell'obiettivo, ovvero quello della navigazione, in diversi sotto-problemi, richiamando quindi delle sub-routine adeguate a risolvere il sotto-problema in questione, arrivando infine auspicabilmente a soddisfare il risultato atteso dall'obiettivo maggiore. Possiamo pensare a questo processo di risoluzione del problema in SHAKEY come una sorta di generatore di piani da seguire: STRIPS utilizza i) ciò è rappresentato nel suo programma attraverso simboli, ovvero che il robot sa riguardo al suo mondo, ii) le possibilità di azione aperte da una determinata condizione dell'hardware nel mondo (anche questo rappresentato simbolicamente) e iii) le routine che è possibile avviare per risolvere il problema per generare una serie di istruzioni che, se eseguite in modo congruo, comportano il raggiungimento dello spazio della soluzione.

Si ricorderà che tale terminologia è stata utilizzata descrivendo il funzionamento della ricerca euristica elaborato da Newell e Simon (cfr. *infra* §2.1.1.2): infatti, il meccanismo di ricerca nello spazio del problema utilizzato da STRIPS, concettualizzato spesso come una ricerca ad albero (*tree search*), ha come riferimenti e precursori proprio i risolutori di problemi come il Logic Theorist e il General Problem Solver. Come sottolineato da Nielssen, «STRIPS operates in a GPS-like manner» (1984: 58); Kuipers e colleghi dichiarano che STRIPS incorpora esplicitamente l'analisi mezzi-fini (*means-end analysis*, Kuipers et al. 2017: 93).

In SHAKEY, questo era stato reso possibile da una riconcettualizzazione del problema della navigazione. Infatti, se visto in un certo modo, il problema di navigazione di SHAKEY era un problema di ricerca in uno spazio del problema, molto simile a quelli già affrontati da programmi di risoluzione di problemi matematici e di dimostrazione di teoremi come il Logic Theorist e il General Problem Solver sopra menzionati. Quello che Nilsson fece, insieme a Bertram Raphael e Peter Hart, fu pensare a una euristica sufficientemente complessa per consentire a SHAKEY di affrontare il problema della navigazione come un percorso ad albero disseminato di nodi, ovvero dei punti cruciali in cui scegliere dove andare. Insieme, i tre ricercatori arrivarono alla elaborazione dell'algoritmo A\* (Hart et al. 1968), probabilmente una delle più importanti invenzioni del gruppo di Stanford e tutt'oggi usato in diverse applicazioni della *computer science*, come il NLP (Russell e Norvig 2010: 93-96). L'algoritmo A\* garantiva a SHAKEY di intraprendere sempre il percorso più breve tra quelli possibili nel suo mondo, tenendo inoltre in considerazione le caratteristiche dell'ambiente, e dimostrava una buona robustezza anche in situazioni di difficoltà (Nilsson 2010: 220). Ciò era dovuto

al fatto che l'algoritmo di navigazione occupava il livello intermedio (ILAs) dell'architettura di SHAKEY, in cui i programmi erano predisposti per raggiungere un determinato obiettivo e continuavano a provare anche se la routine di azione in corso falliva, consentendo al sistema di aggiornare la propria rappresentazione del mondo e di ricalibrare il piano a partire da nuove condizioni. Ad esempio, se il compito di SHAKEY era quello di spingere un oggetto in avanti (PUSH) e l'oggetto sfuggiva dalla sua area di intervento, il programma si accorgeva del problema grazie ai sensori tattili e provvedeva a istruire SHAKEY sul suo riposizionamento, affinché fosse possibile portare a termine il compito (Ibidem: 221).

Nonostante un sofisticato meccanismo di pianificazione ed esecuzione, tuttavia, non era scontato che SHAKEY riuscisse a portare a termine il proprio compito e raggiungere l'obiettivo. A volte l'obiettivo può non essere raggiungibile a causa di impedimenti fisici lungo il cammino, altre volte la rappresentazione del mondo può non essere corretta, spesso i ricercatori spostavano volontariamente dei blocchi o delle scatole nel mondo di SHAKEY per testare le sue capacità di 'reagire' e aggiornare il modello del mondo in cui si trovava per poter portare a termine il compito (spoiler: non spesso ci riusciva).

Per cercare di assistere SHAKEY nel processo di esecuzione e di monitoraggio delle proprie azioni, i ricercatori hanno implementato il programma PLANEX. PLANEX è considerato il livello più alto dell'architettura gerarchica di SHAKEY, che gli consente di monitorare l'esecuzione dei programmi negli strati sottostanti e di riconfigurare la struttura del piano elaborato da STRIPS all'occorrenza. Ad esempio, PLANEX contiene un algoritmo per calcolare il grado di errore in un certo momento di esecuzione del task, considerando la necessità costante della riduzione del 'rumore' nell'esecuzione delle ILAs. L'algoritmo, raggiunta una certa soglia, rende possibile a PLANEX di avvertire SHAKEY che c'è bisogno di fare un controllo sulla rappresentazione del mondo o sulla sua posizione. Se si riconosce un errore significativo, allora PLANEX corregge il piano in accordo alle nuove precondizioni registrate<sup>55</sup>. In sintesi,

Once a plan to accomplish a goal has been constructed, the robot executive system, PLANEX, executes it. If problems arise during execution, PLANEX must also decide how to modify the plan it is executing or whether to construct a new plan. (Nielssen 1984: 65)

La combinazione del planner STRIPS, del programma di esecuzione PLANEX e dell'algoritmo A\* ha reso possibile a SHAKEY di realizzare in modo intelligente le prime forme di comportamento autonomo per un robot in un ambiente controllato. Ma quello che è ancor più rilevante in questa sede è che SHAKEY costituisce un esempio dell'ipotesi del sistema fisico di simboli in azione. Infatti, le

---

<sup>55</sup> L'esempio è preso da Bermudez (2020: 414).

strutture fisiche di simboli si ritrovano in SHAKEY in formule ben strutturate nel calcolo dei predicati, i quali gli forniscono il modello dell'ambiente, i suoi obiettivi e i sotto-obiettivi. STRIPS e PLANEX, poi, sono i programmi attraverso cui le strutture fisiche di simboli vengono trasformate e manipolate, costituendo dunque un sistema intelligente a tutti gli effetti se visto dal punto di vista dell'ipotesi di Newell e Simon. Ma non solo: SHAKEY è anche un esempio dell'ipotesi della ricerca euristica, secondo cui il problem solving intelligente prende forma attraverso la trasformazione e la manipolazione di strutture di simboli fino a che una soluzione pianificata e gerarchizzata non sia raggiunta. Il punto di partenza è dato a SHAKEY dal suo modello dell'ambiente, insieme alla sua formula target che rappresenta lo stato finale desiderato. Attraverso il processo di manipolazione e trasformazione dei simboli eseguito da STRIPS e PLANEX, il problema è risolto quando la struttura di simboli iniziale è stata trasformata nella struttura di simboli da cui la formula target può essere dedotta (Bermudez 2020: 414).

Ora, è possibile considerare la realizzazione di SHAKEY come il primo passaggio di quel movimento elicoidale tipico del metodo sintetico che abbiamo descritto in §1.2.3. Infatti, la realizzazione di SHAKEY coincide con il movimento deduttivo in cui ipotesi teoriche riguardo al funzionamento di meccanismi o funzioni cognitive convogliate nella realizzazione empirica in una macchina: come abbiamo mostrato, SHAKEY è stato costruito seguendo l'ipotesi del sistema fisico di simboli, con i suoi concetti di manipolazione delle strutture e di rappresentazioni simboliche, in congiunzione con le formulazioni del concetto di intelligenza come problem-solving portato avanti tramite ricerche euristiche. È arrivato il momento, dunque, di osservare la macchina costruita che incorpora tali ipotesi, analizzando se le teorie elaborate risultano appropriate a descrivere e/o spiegare i meccanismi oggetto di indagine in prima istanza. Tale analisi è possibile prendendo in considerazione il correlato sull'intelligenza sviluppato da Newell e Simon riguardo l'ipotesi del sistema fisico di simboli. Seguendo il ragionamento, l'ipotesi comporta che ogni macchina capace di manipolare un sistema fisico di simboli debba essere identificata come una macchina intelligente, a prescindere che questa macchina sia una macchina artificiale o biologica (Newell e Simon 1975). Ora, le condizioni necessarie e sufficienti di cui abbiamo parlato in §2.1.1.2 risultano rispettate in SHAKEY? Per quanto riguarda la condizione necessaria, il robot riesce a manipolare la struttura di simboli, per cui può, in potenza, portare avanti azioni intelligenti; per quanto riguarda la condizione sufficiente, d'altro canto, il giudizio non è così immediato. Infatti, la condizione sufficiente afferma che qualsiasi sistema fisico di simboli di una grandezza sufficiente può essere organizzato in modo tale da produrre azioni intelligenti. La domanda, quindi, diventa: riesce SHAKEY a produrre azioni intelligenti?

Se prendiamo in considerazione il parametro impostato da Newell e Simon, in cui l'intelligenza è una capacità di problem-solving informata da processi euristici di ricerca, allora SHAKEY senza dubbio le compie. Ma se paragoniamo le prestazioni registrate da SHAKEY nei compiti a lui assegnati e quelle che potrebbe realizzare un essere umano nello stesso contesto, allora siamo molto lontani dal poterle definire intelligenti. Questo perché, nonostante la straordinaria tecnologia implementata nel robot e le impressionanti abilità mai dimostrate prima, le sue prestazioni erano davvero scarse in termini di emulazione della stessa azione svolta da un essere umano: per attraversare un'ambiente, SHAKEY impiegava diverse ore, fermandosi ogni pochi metri per ricalibrare le sue rappresentazioni e aggiornare il modello di mondo su cui doveva computare i propri ragionamenti; al minimo errore per cui non era prevista la soluzione, SHAKEY si arrestava e poteva solamente essere riportato al punto di partenza; SHAKEY non era capace di percepire alcunché al di fuori di quello che era già rappresentato nel suo mondo e con cui era stato programmato a interagire. Questa breve disamina risponde anche alla domanda: riesce SHAKEY a portare avanti azioni in tempo reale, come era stato per SHRDLU e come volevano i suoi sviluppatori? La realtà dei fatti era che, nonostante gli enormi passi avanti fatti con la realizzazione di SHAKEY, il robot non riusciva affatto ad eseguire azioni in tempo reale, né tanto meno a adattarsi al suo ambiente o a improvvisare nello svolgimento dei processi. I problemi che Winograd aveva sottolineato per SHRDLU si complicano ulteriormente in SHAKEY, che deve aver a che fare non con un mondo virtuale in cui tutti i parametri sono conosciuti e nessun imprevisto può provenire dall'ambiente, ma con uno scenario materiale e – anche se in minima misura – caotico e imprevedibile. In sostanza, nessuna delle prestazioni svolte da SHAKEY poteva far pensare a un'equiparazione con il livello di intelligenza mostrato dagli umani in situazioni simili.

## **2.2 Realizzare un'idea: teoria e pratica della robotica**

Le critiche a SHAKEY sono state tante quante gli elogi. In questa sezione, ci concentreremo principalmente sulle critiche che hanno portato a una separazione dei percorsi tra robotica e intelligenza artificiale. Infatti, a partire dalle riflessioni sviluppate riguardo le prestazioni e l'architettura del robot costruito alla SRI, alcuni interpreti hanno segnalato, dall'inizio degli anni '70 in poi, una biforcazione dei percorsi di ricerca, uno legato all'intelligenza artificiale, l'altro alla robotica (anche detta *robotica cognitiva*, Carlucci Aiello et al. 2012; Cangelosi e Asada 2022).

Per capire i motivi di tale scissione, è utile riflettere ancora sulle implicazioni contenute nella condizione sufficiente dell'ipotesi del sistema fisico di simboli. Come più volte esposto, la condizione sufficiente per l'intelligenza nell'ipotesi del sistema fisico di simboli consisteva nell'avere un sistema «of sufficient size» (Newell e Simon 1976: 293) che sia in grado di manipolare strutture simboliche

per avere come output delle azioni intelligenti. Le implicazioni pratiche che sono derivate dall'interpretazione di questa condizione sono, come dicevamo, almeno due:

- i) Da una parte, il sistema fisico di simboli, nel caso di SHAKEY, ad esempio, potrebbe non essere stato di taglia, grandezza, o complessità sufficiente per produrre azioni e comportamenti intelligenti. Per quelli che hanno interpretato così il problema, la soluzione che si prospettava era quella di continuare a lavorare affinché il sistema si potenziasse, arrivando ad assumere le proporzioni sufficienti per raggiungere il risultato atteso. Tale prospettiva, adottata da pionieri dell'intelligenza artificiale come McCarthy, Minsky (1982), Papert, e gli stessi Newell e Simon, confluiva nell'impresa di aumentare il potere computazionale delle intelligenze artificiali, proponendo una visione del problema in cui il sistema non era ancora in grado di manipolare la giusta *quantità* di dati. In sostanza, quello di cui ancora mancavano i sistemi artificiali era l'abilità di tenere conto di una quantità di dati sufficiente, comparata a quella che riesce a manipolare un sistema (o cervello) umano, per generare l'intelligenza. Si può dire che tale approccio non ha esaurito ancora la sua forza, contribuendo tuttora a sviluppare computer sempre più potenti in grado di gestire quantità di dati sempre più ampie. Questa linea di indagine è quella senza dubbio più approfondita nell'ambito della storia dell'intelligenza artificiale (Davis 2000; Boden 2006).
- ii) Dall'altra parte, una serie sempre più nutrita di studiosi ipotizza che questa non sia la giusta formulazione per una teoria dell'intelligenza. Non importa quanti dati la macchina riuscirà nel corso dell'evoluzione della tecnologia e della disciplina a manipolare: una macchina non potrà mai raggiungere prestazioni intelligenti al livello di quelle umane se si continua sulla strada tracciata dalla GOFAI. Tale linea critica nasce già intorno all'inizio degli anni '50, con pensatori come Bar-Hillel e Taube (Boden 2006: 840-844), ma si sviluppa principalmente a partire dagli anni '70 e acquista un'importanza ancora maggiore visti gli sviluppi eccezionali a cui avrebbe portato nel campo della robotica (cfr. §2.3.2). Le ragioni che portarono alla formazione di questa nuova corrente furono diverse e comprendono riflessioni provenienti da campi del sapere anche molto distanti tra loro, come l'approccio fenomenologico in filosofia, la biologia teorica e studi paralleli nelle scienze cognitive, ma tutte si sviluppano a partire dalla presa d'atto della problematicità delle prestazioni dimostrate dalle macchine computazionali, robot compresi, che erano state costruite fino a quel momento. La linea critica può essere riassunta così: se le macchine che sono state realizzate secondo i principi sviluppati dalla GOFAI non riescono a restituire prestazioni intelligenti al pari o comparabili a quelle degli esseri umani, allora la teoria che supporta

la costruzione di queste macchine deve essere riformulata. La diversità dei campi disciplinari coinvolti in tale riformulazione fa sì che le riflessioni siano variegata e dirette a diversi ambiti dell'intelligenza artificiale e della robotica.

Dunque, la prima linea di ricerca, quella dedicata allo sviluppo della capacità computazionale delle macchine, appartiene alla storia dell'intelligenza artificiale, più che a quella della robotica. È innegabile che notevoli contributi in questo senso siano serviti anche alla robotica per rinsaldare e sviluppare il suo terreno di ricerca, come accaduto ad esempio nel caso della quinta generazione di computer che prende piede in Giappone e che confluisce nella fruttuosa linea di ricerca sui robot umanoidi in cui oggi il Giappone è leader (Togai 1984; Feigenbaum & Shrobe 1993; Akhtaruzzaman & Shafie 2010). Tuttavia, risultati come quelli oggi esibiti da robot umanoidi o androidi non avrebbero potuto realizzarsi senza il contributo di una rivoluzione che prende piede proprio a partire da una critica severa ai presupposti teorici delle scienze cognitive classiche e alle realizzazioni pratiche della GOFAI. Soprattutto, è proprio a partire da alcune severe critiche a SHAKEY che emerge un nuovo modo di vedere la realizzazione di artefatti intelligenti, il quale presuppone strutturalmente non più solamente la costruzione di intelligenze artificiali le cui prestazioni avvengono in mondi virtuali, ma di macchine capaci di interagire con un ambiente, con un'intelligenza *embodied*.

Per capire meglio le ragioni di questa parte, è necessario riportare alcune delle critiche sviluppate a SHAKEY da uno dei pionieri della *Nouvelle AI*, Rodney Brooks. Analizzeremo nel dettaglio le idee sviluppate da Brooks per una nuova robotica in §2.3.2, ma è utile partire da qui perché è proprio Brooks a notare come sia a seguito della realizzazione di SHAKEY che il panorama dell'intelligenza artificiale ha cominciato a mostrarsi debole, coinvolgendo tutti gli ambiti della sua applicazione.

The Shakey project is often quoted as a smashing success of applied AI. In contrast I feel its methodology was misguided, and caused a misdirection of the bulk of AI research during the seventies and eighties. AI is only just now beginning to recover from the disaster. (Brooks 1989: 445).

In diversi articoli, Brooks (1989, 1991a, 1991b, 2002) sviluppa una dettagliata riflessione sugli aspetti problematici di SHAKEY e sulle possibili cause di quello che definisce un fallimento per l'intelligenza artificiale. Il primo fattore è riconducibile all'assunto implicito portato avanti dalla corrente computazionalista delle scienze cognitive riguardo la possibilità di separare l'intelligenza dai processi percettivi e dall'azione, considerandola una capacità astratta di *problem-solving* (cfr. Newell e Simon 1975; Minsky 1961). Nel suo articolo "The Whole Iguana" (1989), Brooks argomenta che effettivamente SHAKEY sembrava validare l'ipotesi che un robot potesse davvero operare nel mondo pur avendo moduli separati per percezione, azioni e intelligenza (che risiedeva nei sistemi di pianificazione e controllo). Tuttavia, secondo Brooks, «Shakey cheated» (Ibid.: 446),

perché, come abbiamo visto, il suo mondo era appositamente creato per lui, costruito come una semplificazione spogliata dagli imprevisti e dalle eccezioni che sono invece tipiche del mondo ‘naturale’. Le ragioni per la semplificazione del mondo sono riconosciute da Brooks dipendere anche dallo stato di avanzamento delle tecnologie dell’epoca – come, ad esempio, l’impossibilità di computare input visivi – ma resta il fatto che il risultato conduceva a una simulazione non veritiera dell’ambiente. In questa caratteristica risiede il secondo fattore problematico incarnato da SHAKEY, ovvero il fatto che il mondo per lui elaborato fosse rappresentato nella sua interezza simbolicamente all’interno dell’architettura del robot. La pratica di mappatura dell’ambiente nell’architettura del robot sembrava dimostrare che una rappresentazione *adeguata* del mondo fornita all’agente era tutto quello di cui necessitava per poter agire efficacemente nel mondo fisico: nessuna importanza veniva data ai vincoli materiali o all’imprevedibilità di un ambiente reale.

Da qui, Brooks fa derivare due importanti riflessioni circa gli inconvenienti incontrati dal robot della SRI. In primo luogo, l’assunzione implicita che ci sia la possibilità di rappresentare perfettamente il mondo giocattolo conduce a considerare erroneamente possibile il fatto che questo possa accadere anche negli agenti biologici con il mondo ‘reale’; in secondo luogo, il fatto che il mondo sia ‘chiuso’ comporta anche che il robot possa sapere tutto ciò che accade al suo interno, fattore di impossibile realizzazione negli scenari non controllati o non costruiti. Il fatto che i ricercatori abbiano posto la loro fiducia sul modello dell’intelligenza sviluppato dalla GOF AI ha comportato l’impresa di costruire un robot che rappresentasse totalmente il mondo per muoversi al suo interno, come diretta conseguenza dello sviluppo della teoria. D’altronde, SHRDLU sapeva muoversi all’interno di un mondo virtuale ed eseguiva i comandi con sufficiente accuratezza: «if a program could be built which understood such a world completely and could also manipulate that world, then it was assumed that programs which assumed that abstraction could in fact be connected to the real world without great difficulty» (Brooks 1991a: 9). In teoria, niente faceva immaginare che la costruzione di un robot secondo i principi della GOF AI come era SHAKEY – anche definito ‘un computer con le ruote’ – avrebbe portato a mettere in discussione tutto l’impianto teorico che lo aveva ispirato. In pratica, questo fu proprio quello che accadde, e costituì uno dei fattori che contribuì a far cadere non solo la robotica, ma anche molte delle ricerche nel settore dell’intelligenza artificiale, in un profondo inverno (*AI Winter*; cfr. Umbrello 2021).

Aggiungendosi alle critiche di Brooks, altri notarono che le difficoltà dimostrate da SHAKEY potessero essere imputate alle ipotesi riguardo al funzionamento dell’intelligenza che il robot incorporava. Ad esempio, Hans Moravec (il quale era stato studente a Stanford mentre i report di SHAKEY erano in preparazione) sottolinea che nessuna delle sequenze che prevedevano azioni, come andare in una stanza, trovare un blocco, spostare il blocco nella scatola o spostare la scatola,



era mai stata eseguita dal robot senza soluzione di continuità (Boden 2006: 867). Husbands (2014: 275) sottolinea che il dispendio computazionale a cui il robot era sottoposto nel processo di esecuzione e monitoraggio dell'attività di problem-solving portato avanti da PLANEX era uno dei fondamentali fattori di rallentamento delle performance di SHAKEY, che inficiava non di poco la qualità e le tempistiche di realizzazione delle sue azioni. Il processo inferenziale di manipolazione di simboli, unito alla bassa qualità degli algoritmi di decifrazione di visione e all'enorme lavoro di aggiornamento delle rappresentazioni interne, portavano SHAKEY a dover compiere un lavoro colossale anche solo per spostarsi di un metro (come sarebbe successo anche per il robot Stanford Cart, cfr. Moravec 1988).

Non mancarono dei sottili attacchi anche da parte di esponenti dello stesso approccio GOFAI alle scienze cognitive. In un'intervista condotta da Jeremy Bernstein, Minsky confida di essere stato scettico nei confronti di SHAKEY fin dagli inizi.

Pensavo che fosse già problematico cercar di comprendere gli occhi, le mani, ecc.; perché aggiungere inutilmente anche problemi di ingegneria? Contro il mio parere, gli amici dello Stanford Research Institute decisero di dotare di movimento il loro primo robot: lo chiamarono Shakey (Tremulo), e lo era proprio. Procedeva oscillando e traballando, fino a strapparsi letteralmente i cavi, e le continue vibrazioni impedivano alla telecamera di fare buone riprese. Inoltre richiedeva un gran lavoro di manutenzione, e credo che già questo impegno abbia impedito loro di concentrarsi su altri aspetti. (Bernstein 1978/1982: 154)

In sostanza, è attestato che il grado di innovazione apportato da SHAKEY, se nel campo della tecnologia dell'intelligenza artificiale ha contribuito notevolmente allo sviluppo di sistemi avanzati per il controllo e il problem-solving, dal punto di vista della teoria e delle formulazioni nelle neonate scienze cognitive poneva numerosi interrogativi.

In questo viaggio di andata abbiamo analizzato come il processo deduttivo che costituisce il primo movimento del metodo sintetico abbia avuto in SHAKEY la sua realizzazione pratica. Le teorie del sistema fisico di simboli e delle ricerche euristiche sono state implementate in una macchina, l'osservazione del comportamento della quale può comportare, all'interno delle scienze cognitive, sia la validazione dell'ipotesi di partenza, sia una sua ridiscussione. Nel primo caso, la linea di ricerca che ha considerato SHAKEY come una (parziale) validazione delle teorie implementate prende una sua propria strada, proseguendo nella direzione di quella che sarebbe diventata la storia ufficiale dell'intelligenza artificiale (Boden 2006); nel secondo caso, le riflessioni innescate dalla messa in discussione dei principi su cui SHAKEY era stato costruito avranno un impatto rivoluzionario sulla robotica degli anni '80 e '90. Questa seconda strada è quella che ci preme indagare ora.

Il prossimo paragrafo, quindi, sarà dedicato a un viaggio di ritorno: ritorno della pratica sulla teoria – dalla robotica alle scienze cognitive, un viaggio che sarà disseminato di riflessioni riguardo al modo in cui sia possibile creare artefatti intelligenti che si basano su principi alternativi a quelli della

GOFAI, in cui si è messa in moto una riforma delle idee riguardo al funzionamento della mente e della cognizione, al ruolo del corpo, al legame tra percezione e cognizione, alla natura dell'intelligenza, ripensando presupposti dati per assodati nell'indagine dell'intelligenza umana e artificiale.

## **2.3 Viaggio di ritorno: dalla robotica alle scienze cognitive**

### *2.3.1 Contra GOFAI*

Come abbiamo visto, le critiche alla GOFAI non si esauriscono con gli attacchi alla sola componente ingegneristica di SHAKEY. Anche, e soprattutto, le ipotesi che avevano nutrito la sua realizzazione sono state oggetto di attacchi mirati da parte di filosofi e scienziati cognitivi. Tra i più celebri, a sferrare la critica più acuta (e più longeva, visto che continua a essere efficace ancora oggi) è quella elaborata dal filosofo John Searle (1980; Boden 2006: 1382-1385). Gli argomenti di Searle si sviluppano come una risposta all'idea che un sistema capace di manipolare simboli possa essere effettivamente un sistema intelligente, affermando invece che tutto quello che è prodotto dalla macchina computazionale sia "all syntax and no semantics", anche se il risultato prodotto è corretto. In sostanza, la risposta di Searle all'ipotesi del sistema fisico di simboli è che ciò che la macchina fa consiste nel manipolare simboli secondo le regole che gli vengono impartite, senza di fatto comprendere alcunché né dei processi di computazione che porta avanti, né del contenuto semantico, ovvero il significato, che questi processi producono come output. Per dimostrarlo, Searle elabora l'esperimento mentale poi diventato celebre come l'argomento della stanza cinese (Searle 1980: 417-419), in cui il filosofo chiede di immaginare una stanza chiusa in cui c'è una persona che riceve da sotto la porta foglietti di carta e passa a sua volta dei foglietti di carta fuori dalla stanza attraverso un'altra porta. Sui foglietti in entrata sono riportati dei simboli in cinese, ma la persona all'interno della stanza non conosce questa lingua. Tuttavia, grazie a un manuale di istruzioni che associa a ogni simbolo scritto sul foglietto in entrata un altro simbolo da dare come risposta nel foglietto in uscita, la persona riesce comunque a passare il giusto foglietto all'esterno della stanza, dando l'impressione a chi raccoglie i foglietti da fuori che chi c'è nella stanza capisca perfettamente il cinese. Ma in realtà la persona all'interno della stanza non capisce cosa sta dicendo, se sta dichiarando guerra a uno stato nemico, proponendo un matrimonio a un suo spasimante, o elaborando argomenti filosofici riguardo la possibilità di un AI senziente. Tutto ciò che viene fatto dalla macchina, dice Searle, è completamente scollegato dai meccanismi di generazione o attribuzione di significato che gli esseri umani realizzano quando danno senso al mondo:

[...] in the literal sense the programmed computer understands what the car and the adding machine understand, namely, exactly nothing. The computer understanding is not

just (like my understanding of German) partial or incomplete; it is zero. (Searle 1980: 419).

Questa critica ha innescato un gran fermento nel dominio della filosofia applicata all'intelligenza artificiale, producendo anche numerose risposte. Quella che ci interessa osservare in questa sede è quella che è stata definita dallo stesso Searle come "Robot reply":

Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking – anything you like. The robot would, for example, have a television camera attached to it that enabled it to 'see', it would have arms and legs that enabled it to 'act', and all of this would be controlled by its computer 'brain'. Such a robot would [...] have genuine understanding and other mental states. (Searle 1980: 420)

Secondo questa obiezione mossa all'argomento della stanza cinese di Searle, un robot come SHAKEY dovrebbe avere la capacità di comprendere realmente ciò che accade nel proprio mondo, essendo questo rappresentato nella sua 'mente' e potendo agire fisicamente al suo interno. In controposta, Searle nota come questa obiezione sia essa stessa una variazione controproducente per l'ipotesi che l'intelligenza sia il risultato di una manipolazione di simboli, in quanto aggiunge all'attività di manipolazione stessa anche azioni che comprendono l'interazione con il mondo (come il mangiare e il bere). E comunque, sottolinea il filosofo, il discorso riguardo all'incapacità di comprendere della macchina non cambia: anche se il robot fosse capace di rintracciare da solo i 'simboli' da manipolare nell'ambiente, eliminando il contributo esterno dell'attività proveniente dagli input in entrata, il processo di manipolazione resta lo stesso; anche se i simboli si riferiscono alle attività reali del mondo, il robot non acquista la capacità di renderle significative automaticamente: per Searle, c'è sempre qualcosa che manca, che per lui è l'intenzionalità in quanto fenomeno biologico<sup>56</sup> (Searle 1980; cfr. §4). L'argomento della stanza cinese di Searle, e il richiamo a una componente imprescindibile dell'intelligenza biologica, sembra non solo porre un freno ai giudizi impulsivi di quelli che credevano che sarebbe stato possibile creare macchine intelligenti al pari dell'essere umano in una manciata di decenni (come, ad esempio, Herbert Simon o Marvin Minsky, cfr. Boden 2006), ma si configurava come una dimostrazione teorica dell'inadeguatezza delle teorie sviluppate fino a quel momento nelle scienze cognitive legate all'artificiale.

Il discorso di Searle si pone agli antipodi rispetto alle posizioni classiche riguardo mente, intelligenza e il suo possibile surrogato artificiale, ma si basava su una serie di riflessioni che avevano

---

<sup>56</sup> «Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origin as lactation, photosynthesis, or any other biological phenomenon. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not». (Searle 1980: 424)

iniziato a svilupparsi già a partire dagli anni '70. In particolare, un gruppo di ricercatori attivi nel campo della filosofia dell'AI e delle scienze cognitive si era reso conto che i risultati raggiunti fino a quel momento nel merito della replicazione di facoltà cognitive di alto livello, come il *problem-solving*, l'algebra e la matematica tutta o il ragionamento strategico, era stata molto più 'facile' rispetto a problemi che non si riteneva importante risolvere per ricreare un comportamento intelligente. Attività come camminare, afferrare materialmente oggetti, vedere e dare senso al mondo in modo spontaneo erano considerate operazioni del tutto distaccate dal ragionamento intelligente, considerato di norma risiedere in ragionamenti astratti con un'alta componente simbolica. Di fatto, si pensava che l'insufficienza delle prestazioni dimostrate dai robot rispetto a quelle eccellenti che man mano venivano performati dagli altri sistemi artificiali potesse essere imputabile alle incapacità di implementazione dei ricercatori in robotica e non ai difetti delle ipotesi. A suggerire esplicitamente che questa non fosse l'interpretazione corretta dei fatti, ma che il problema risiedesse altrove, fu proprio Hans Moravec:

It seemed to me, in the early 1970s, that some of the creators of successful reasoning programs suspected that the poor performance in the robotics work somehow reflected the intellectual abilities of those attracted to that side of the research. Such intellectual snobbery is not unheard of, for instance between theorists and experimentalists in physics. But as the number of demonstrations has mounted, it has become clear that it is comparatively easy to make computers exhibit adult-level performance in solving problems on intelligent tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility. (Moravec 1999: 15)

Insomma, allenare un computer a diventare un campione di scacchi è fattibile, ma l'importante è non pretendere che sia lui stesso a muovere i pezzi della scacchiera. Questo argomento diventò presto celebre con il nome di *paradosso di Moravec*, denotando proprio la difficoltà di implementare nei sistemi di AI quelle capacità cognitive considerate più 'semplici', ovvero quelle date per scontate quando si parla di intelligenza. Moravec imputò tale paradosso al fatto che, se le funzioni cognitive di alto livello erano state sviluppate nel corso dell'evoluzione in poco meno di un centinaio di migliaia di anni, i meccanismi basilari attraverso cui un agente si comporta in modo intelligente, come la percezione visiva e tattile o la capacità di agire efficacemente nell'ambiente, sono costati all'evoluzione milioni di anni. Il paradosso consiste nel presupporre che più l'attività abbia richiesto tempo all'evoluzione, più tempo ci vorrà per implementare quella stessa funzione in artefatti prodotti dall'essere umano, anche se a prima vista potrebbero sembrare attività che richiedono uno sforzo cognitivo di gran lunga inferiore rispetto a compiti come il pensiero strategico o la pianificazione. Scoraggiante per chi credeva di poter raggiungere risultati strabilianti nel giro di qualche decennio, ma comunque possibile: in modo differente da Searle, Moravec credeva che fosse possibile replicare qualche funzione cognitiva appartenente all'intelletto umano, a patto che si imiti l'evoluzione della

mente animale «by striving to add capabilities to machines a few at a time, so that the resulting sequence of machine behavior resembles the capabilities of animals with increasingly complex nervous system» (Ibid: 17).

La linea argomentativa sviluppata da Moravec ci suggerisce come la robotica abbia giocato un ruolo fondamentale di ancoramento della teoria alla pratica di realizzazione degli artefatti artificiali, aggiungendo gradi di complessità derivanti dal rapporto materiale con l'ambiente e dall'interazione diretta in tempo reale che nelle applicazioni dell'intelligenza artificiale risultavano ancora non così evidenti. Questioni quali il *frame problem* (McCarthy e Hayes 1968; Shananan 2016; cfr. §2.3.3), ad esempio, nella robotica diventano ancora più evidenti (Agre 1985; Dennett 1990; Wheeler 2008), costituendo una prova ulteriore del fatto che, per costruire un'intelligenza artificiale, robotica o meno, caratteristiche date spesso per scontate come la corporeità dovevano essere incluse nel quadro teorico sulle ipotesi del funzionamento della mente.

In sostanza, le lacune dimostrate dalla GOFAI nella realizzazione delle sue ipotesi vengono bersagliate da diversi punti di vista. In particolare, le critiche che porteranno all'elaborazione di un nuovo approccio in robotica possono essere raccolte in tre argomenti cardine: i) la metafora del computer per descrivere la cognizione umana è una chimera che ha condotto a numerosi errori di considerazione sulla natura della mente; ii) la logica formale non è il giusto strumento a cui ancorare la realizzazione dell'intelligenza artificiale se l'obiettivo è quello di replicare i meccanismi cognitivi dell'essere umano; iii) l'intelligenza artificiale non sarà realizzabile fino a che non verrà concettualizzata come risultato delle azioni di un corpo capace di esperire e dare significato autonomamente.

Questi tre argomenti hanno, nei fatti, prodotto un'autonomizzazione dei tre domini che, fino a SHAKEY, avevano proceduto di pari passo e in perfetta sinergia: l'intelligenza artificiale, la robotica e le scienze cognitive. Il risultato è stata una riconfigurazione dei rapporti tra questi domini e una moltiplicazione delle loro possibili interazioni. Per questo, nei prossimi paragrafi esporremo le caratteristiche principali di ognuna di queste tre argomentazioni fondamentali, mostrando infine i loro risultati sull'episteme robotica.

### *2.3.1.1 Non farti ingannare dalle metafore: Joseph Weizenbaum*

Abbiamo già trattato il tema della metafora in diversi momenti: abbiamo analizzato in che modo lo studio della cognizione e del comportamento umano attraverso macchine abbia condotto alla meccanizzazione delle indagini che hanno come oggetto l'umano sotto vari rispetti, dalla medicina alla filosofia, dalla biologia alle scienze della mente, e ci siamo concentrati su una metafora specifica, quella che costruisce una proporzione strettissima tra mente e software da una parte e cervello e

hardware dall'altra, ovvero quella della cognizione come computer adottata e sviluppata dall'approccio computazionalista delle scienze cognitive della GOF AI. Fin dalla pubblicazione da parte di von Neumann della sua monografia *The Computer and the Brain* (1958), la metafora computazionale viene posta al centro dell'indagine dei processi cognitivi.

I primi segnali di malcontento rispetto all'uso pervasivo della metafora computazionale per le scienze dell'artificiale (scienze cognitive comprese) si cominciano a presentare già a partire dagli anni '60, con un movimento di contro-tendenza rispetto alle ricerche allora in corso sull'intelligenza artificiale che si riaggancia anche a tradizioni precedenti rispetto a quelle computazionaliste, come la cibernetica (Boden 2006: 852-855). Tuttavia, le critiche più serrate e focalizzate cominciano a svilupparsi a partire dall'inizio degli anni '70, con interpreti come Mortimer Taube, Yehoshua Bar-Hillel (Boden 2006: 846-848) e Hubert e Stuart Dreyfus (Dreyfus 1965, 1967, 1972, 1992; Dreyfus e Dreyfus 1986a, 1991). Inoltre, il critico forse più inaspettato, data la sua particolare posizione di sviluppatore di uno dei programmi più di successo della GOF AI, è Joseph Weizenbaum. Celebre per aver inventato ELIZA (Weizenbaum 1966; cfr. §2.1.1.3), il già menzionato programma che consentiva a qualsiasi persona di dialogare con una riproduzione di uno psichiatra rogeriano, Weizenbaum si distacca lentamente ma inesorabilmente dalle linee di ricerca della GOF AI, individuando nelle interpretazioni dei suoi colleghi delle derive rispetto alle intenzioni iniziali che erano confluite nello sviluppo di ELIZA (Baumgartner e Payr 1996: 249-264). In particolare, ciò che nota Weizenbaum è una tendenza a considerare la metafora computazionale come non più problematica, una forma di comprensione stabilizzata che è entrata a tutti gli effetti a far parte del senso comune:

The computer has become a source of truly powerful and often useful metaphors. Curiously, [...] the public embrace of the computer metaphor rests on only the vaguest understanding of a difficult and complex scientific concept (here, theory of computability and the results of Turing and Church concerning the universality of certain computing schemes). The public vaguely understands – but nonetheless is firmly convinced – that any effective procedure can, in principle, be carried out by a computer. Since man, nature, and even society carry out procedures that are surely “effective” in one way or another, it follows that a computer can at least imitate man, nature, and society in all their procedural aspects. Hence everything (that word again!) is at least potentially understandable in terms of computer models and metaphors. (Weizenbaum 1976: 157).

Secondo Weizenbaum, la catacresizzazione della metafora computazionale aveva condotto a un uso non adeguato delle stesse tecnologie che aveva sviluppato. Infatti, il suo collega Kenneth Colby aveva dichiarato di avere intenzione di usare ELIZA per sostituire sistematicamente l'interlocutore esperto nei dialoghi psichiatrici, utilizzandolo di fatto come un sistema esperto alla pari di DENDRAL o MYCIN, e che questo sarebbe stato l'inizio del processo di automatizzazione della pratica psichiatrica. Weizenbaum si oppose fermamente a tale intento (Boden 2006: 850): da una parte, non

credeva che ELIZA sarebbe stata capace di sostenere conversazioni a livello di quelle che solitamente avvengono tra paziente e psichiatra, denunciando una sfiducia nella tecnologia implementata nella macchina; dall'altra, ciò che maggiormente preoccupava Weizenbaum era che persone con poca familiarità rispetto ai dispositivi di intelligenza artificiale potessero facilmente essere ingannate e portate a credere che dietro alla macchina si nascondesse un qualche esperto capace di elargire consigli e istruzioni informate. Non era, d'altronde, solo una supposizione: era già accaduto che la sua segretaria gli chiedesse di concederle un po' di privacy mentre interagiva con ELIZA<sup>57</sup>. Questo poteva essere dovuto al fatto che, probabilmente, non si vuole compagnia quando si stanno confessando segreti circa i propri traumi infantili, a prescindere dalla natura organica o artificiale dell'interlocutore, come accade ad esempio con un diario segreto; cionondimeno, il pericolo di una sovrainterpretazione delle capacità mostrate da ELIZA restava plausibile, facendo sorgere in Weizenbaum i primi quesiti etici sull'uso dell'intelligenza artificiale in ambito clinico e terapeutico.

Facendo riferimento alle riflessioni scaturite dalla programmazione e dai possibili impieghi di ELIZA, seguendo una corrente già aperta dalle critiche filosofiche di stampo fenomenologico, e tenendosi lontano dai tecnicismi e dal linguaggio astruso degli informatici, Weizenbaum elabora una argomentazione stringente riguardo all'inadeguatezza della metafora computazionale:

Whether or not [AI] program can be realized depends on whether man really is merely a species of the genus "information-processing system" or whether he is more than that. I shall argue that an entirely too simplistic notion of intelligence has dominated both popular and scientific thought, and that this notion is, in part, responsible for permitting artificial intelligence's perverse grand fantasy to grow. I shall argue that an organism is defined, in large part, by the problems it faces. Man faces problems no machine could possibly be made to face. Man is not a machine. I shall argue that, although man most certainly processed information, he does not necessarily process it in the way computers do. Computers and men are not species of the same genus. (Weizenbaum 1976: 203)

Quello che Weizenbaum mette al centro della sua argomentazione critica non è tanto l'uso della metafora computazionale nella sua accezione generale, che nota aver di molto fatto avanzare lo stato della tecnologia e contribuito a portare avanti la conoscenza in molti campi di ricerca, quanto la sua mancanza di messa in discussione e la perdita dello stesso statuto metaforico della macchina nei confronti della mente e della cognizione. Il pericolo presentato da Weizenbaum è che, se continuiamo a non problematizzare e relativizzare la metafora della macchina come mente, si passerà a trattare gli umani come delle macchine.

---

<sup>57</sup> La credenza diffusa che ELIZA possedesse capacità superiori a quelle che effettivamente aveva ha trovato una sua identità nella formulazione dell'*effetto ELIZA* (Hofstadter 1995: 171-186), che denota la fallace credenza che un programma per computer (o un robot) possieda capacità cognitive maggiori o più complesse – più simili all'essere umano – di quanto realmente possibile. Parleremo più diffusamente di questo effetto nei confronti dei robot, anche senza fare ricorso questo nome, in §4.

Possiamo contribuire allora, a problematizzare e mettere in discussione la metafora della mente come macchina. Innanzitutto, va notato che la metafora non solo fornisce uno strumento concettuale molto potente per riuscire a costruire comprensibilità per concetti e oggetti di studio ancora sconosciuti, ma istruisce anche una via maestra per l'indagine degli oggetti indagati. È il caso, senza dubbio, della metafora computazionale che si istaura sulla scorta delle conquiste tecnologiche e che si attesta come linea direttrice nell'indagine della mente, fornendo materiale concettuale per i paragoni e le analogie tra la mente umana e il funzionamento della neonata scienza dell'artificiale. Se una metafora può essere valutata anche in rispetto alla qualità dei prodotti teorici che contribuisce a generare (West e Travis 1991), allora la metafora della cognizione come computer può sicuramente dirsi fruttuosa. Tuttavia, è importante vedere come l'instradamento di una prospettiva di ricerca che segue quasi pedissequamente il *diktat* della metafora in oggetto, oltre a illuminare le possibilità di comprensione e commensurabilità tra i termini in gioco, di fatto oscura le alternative, narcotizzando il pensiero che dalle alternative potrebbe fiorire. Nei termini in cui la pone Weizenbaum, la metafora computazionale, grazie anche alla sua efficacia, comporta che i riflettori siano puntati sulla natura processuale, gerarchica, dualista, macchinica dei processi di pensiero, mettendo in ombra tutto ciò che non rientra sotto questa concettualizzazione e anestetizzando i tentativi di riformulazione secondo metafore magari meno efficaci al momento, ma che potrebbero far fiorire nuovi modelli di concettualizzare la cognizione umana.

In secondo luogo, e in modo complementare alla argomentazione precedente, va notata una caratteristica, che si può definire di strutturale, della metafora e del suo uso nel dominio della scienza: la direzionalità. Come fatto notare da più parti (MacCormac 1985; West e Travis 1991) il movimento della metafora ha una sua importanza: infatti, anche nel dominio linguistico, tra i molti effetti che la metafora produce, questa consente di far vedere in un oggetto di cui poco si conosce delle caratteristiche tipiche di un altro oggetto, che è invece ben conosciuto. In sostanza, tale movimento procede dal conosciuto allo sconosciuto, rendendo possibile allargare il dominio della comprensione laddove prima sembrava dominare l'inintelligibilità<sup>58</sup>. Uno degli esempi classici della direzionalità della metafora nella scienza è la coppia costituita da sistema solare e struttura atomica: ai primi del '900, il primo, con il suo legame gravitazionale noto da secoli, è servito a illuminare la struttura ancora non ben conosciuta della seconda, che invece era ancora di là da comprendere. Alcuni autori (MacCormac 1985; West e Travis 1991) notano che è molto difficile che la direzionalità di una metafora cambi, a meno che il secondo termine, quello che dappprincipio di presentava come

---

<sup>58</sup> È doveroso precisare che non si vuole ridurre *in toto* la struttura della metafora al movimento direzionale che va dal termine conosciuto allo sconosciuto, ma in questa sede tale movimento è selezionato tra i molti altri possibili poiché pone al centro l'uso metaforico di alcuni concetti al fine creare comprensibilità nel dominio della scienza. Per una trattazione complessa della struttura della metafora e degli effetti che questa produce, in modo particolare nel dominio linguistico, si veda Eco (1984).



sconosciuto, non diventi più noto di quello che era stato usato per illuminarlo. Il caso della metafora computazionale sembra essere uno di questi rari casi: la macchina computazionale, dapprima utilizzata per spiegare la struttura della mente umana, è passata a essere spiegata da questa. La macchina come mente era diventata la mente come macchina. Tanto innovativa da sembrare magica (Haugeland 1985: 63-64, 69), la macchina computazionale era diventata la concettualizzazione perfetta della mente e della cognizione umana, tanto da identificarsi saldamente con esse e inaugurare un campo di studi florido e fruttuoso come la psicologia cognitivista o computazionale (Neisser 1967; Marraffa e Paternoster 2012).

Nonostante tale riflessione illumini un aspetto stringente del movimento di evoluzione della metafora computazionale, quello che stabilisce la logica dell'indagine della macchina computazionale attraverso le ipotesi generate per la mente (e non più solo il contrario), è doveroso fare un passo avanti. A ben vedere, infatti, più che un capovolgimento o un'inversione della direzionalità, possiamo parlare di una totale equiparabilità dei termini, chiamati ora a spiegarsi a vicenda. Avevamo già accennato a questo movimento come il terzo passaggio di un percorso storico e filosofico dell'evoluzione della metafora dell'essere umano come macchina (§1.2.4-1.2.5), e qui si fa ancora più chiaro: la mente e la macchina smettono di intessere rapporti metaforici unidirezionali, coniugandosi invece come i termini medi in un anello metaforico che inscena un rapporto di identità. Ma, come fatto notare in un bellissimo passaggio da West e Travis (1991: 69):

Minds are organically based, not programmed, inconsistent, only generally persistent, poor at math, great at pattern recognition, and so on, but computers are the exact opposite in essentially every way. Computers incorporate transistors, consist of clearly defined and limited-function electric circuits, operate on discrete packets of data, maintain memory in specific locations, are generally nonadaptive, are completely deterministic, are brittle (in the sense that power interruptions or magnetic fields render them useless), cannot deal with ambiguity in a satisfactory manner, take immense amounts of time to perform certain classes of operations, and are amazingly accurate. Minds, however, exhibit none of these properties.

Weizenbaum punta allora i riflettori del discorso etico ed epistemologico sull'intelligenza artificiale proprio su questo punto: l'identificazione dei termini a cui si è arrivati tramite lo sviluppo congiunto di scienze cognitive computazionaliste e scienze dell'artificiale ha oscurato ciò che di particolare e di non condiviso, di caratterizzante, era presente in ognuno dei concetti in quanto tale. Il punto è quindi decostruire la metafora computazionale, analizzando il modo in cui le caratteristiche comuni possono essere rese note e sfruttate per costruire comprensibilità, ma allo stesso tempo riscoprirne lo scarto reciproco e salvaguardare la specificità di ognuna delle categorie convocate nella commistione metaforica.

Una conseguenza dell'uso dilagante di tale metafora nelle scienze cognitive computazionaliste è consistita nell'identificare la totalità dei processi di pensiero con la logica formale. Lo abbiamo visto

con la formulazione della teoria della GOFAI e con SHAKEY: anche nel caso in cui fossero state le funzioni non propriamente di alto livello come il ragionamento astratto e il *problem-solving* a dover essere implementate, come la locomozione o la visione, queste venivano comunque riportate a un processo di formalizzazione che le astraeva dal contesto e le riduceva ad operazioni logiche computate internamente. Il movimento successivo, allora, che spetta a una critica della GOFAI è quello di disimplicare la logica formale dalla sua pervasività nelle strutture cognitive, e trovare delle alternative (a loro volta metaforiche) con cui poter sostituire la metafora computazionale che si era dimostrata inadeguata. Negli anni fiorenti dello sviluppo della biologia teorica, con contributi di spicco come quelli di Heinz von Foerster, Jerry Lettvin, Humberto Maturana e Francisco Varela, è allora proprio la biologia che si candida a divenire la disciplina da cui trarre metafore per illuminare concetti inerenti al funzionamento della mente che non potevano essere inclusi nel computo razionalistico della teoria computazionalista.

#### 2.3.1.2 Dalla logica alla biologia: Heinz von Foerster e Terry Winograd

Il motivo per cui, negli anni '70, la biologia poteva candidarsi a disciplina di riferimento delle scienze della mente – e dell'artificiale – risiedeva nella sua più che trentennale commistione con la cibernetica. Anche se la disputa tra cibernetica e intelligenza artificiale era stata “vinta” dalla seconda, che era diventata la disciplina regina nelle scienze dell'artificiale, la cibernetica aveva continuato a sviluppare la sua linea di ricerca intrecciandola saldamente a quella delle scienze della vita. Anche negli anni di relativo silenzio della disciplina, infatti, non sono mancati contributi notevoli all'avanzamento della comprensione dei processi cognitivi, come ad esempio lo sviluppo della teoria dei sistemi di von Bertalanffy (1968) o le idee rivoluzionarie di Gregory Bateson e di Heinz von Foerster. Insieme ad altri contributi, questi daranno vita a quella che è stata definita dallo stesso von Foerster come la “seconda cibernetica” o “cibernetica di secondo ordine” in cui, affondando le proprie radici nelle nozioni sviluppate da Ashby (1962), si sviluppa una riflessione epistemologica non solo in merito ai sistemi complessi e alle loro proprietà organizzazionali (prospettiva che indentificava la cibernetica di primo ordine come disciplina dei sistemi osservati), ma si passa a indagare anche il ruolo dell'osservatore, adottando l'etichetta di disciplina che indaga i *sistemi che osservano* (von Foerster 1974, 1987, 2003; Ceruti 1986, 1989; Damiano 2010; Froese 2010). In particolare, il connubio tra biologia e scienze dell'artificiale attraverso la cibernetica si istituzionalizza proprio con le indagini di von Foerster, fondatore nel 1958 del BCL (*Biological Computer Laboratory*) presso l'Università dell'Illinois. Il suo scopo era quello di indagare concetti considerati il cardine della vita e dell'intelligenza sotto la guida del concetto di *causalità circolare* (Damiano 2010: 65) e di trovarne la specificità nel dominio cognitivo anche attraverso la realizzazione di artefatti artificiali (von

Foerster 2003: vii-ix): realizzare un computer biologico era l'obiettivo principale. Come sottolineato da Ceruti e Damiano (2018), lo scopo del computer biologico era quello di far emergere le caratteristiche della cognizione che venivano oscurate nel modello computazionale:

Living systems, unlike man-made computers, are self-organizing systems, and develop cognition not as abstract problem solving, but as the inherently biological activity of surviving in an ever-changing "hostile" environment. [...] Von Foerster grounded this reformulation in a principle of continuity between life and cognition. Based on this principle, and on a literal understanding of the Latin verb "computare" as "considering things together," he redefined computation as every operation that a living system, as a self-organizing network of processes, is able to perform to preserve itself (von Foerster, 1974). It is a reconceptualization that transfers the notion of computation from the epistemological space of abstract problem solving to that of biological adaptation. (Ceruti e Damiano 2018: 4)

Giocano un ruolo cruciale, allora, gli studi sui meccanismi basilari della vita stessa, tra cui forse il più centrale è il concetto di *auto-organizzazione* (von Foerster 1960), che contribuisce enormemente a cambiare una prospettiva che fino ad allora si era data quasi unilateralmente, ovvero quella tra l'agente (l'organismo) e il mondo. Rispetto ai modelli precedenti o coevi, ciò che von Foerster mette in luce è la strettissima co-dipendenza di organismo e ambiente: seguendo il secondo principio della termodinamica, secondo cui l'entropia<sup>59</sup> non può decrescere in un sistema isolato, l'organismo crea ordine interno a sé stesso attraverso l'assorbimento di disordine dall'esterno (principio dell'ordine dal rumore<sup>60</sup>). Ciò implica che, non potendo il sistema biologico essere isolato, dato che l'entropia non può decrescere, il sistema deve essere inteso non più solamente come struttura del singolo organismo, l'unità biologica individuale, ma sempre e necessariamente in accoppiamento con l'ambiente in cui il sistema è inserito. Solamente così, secondo von Foerster, è possibile pensare all'organismo biologico come potenziale sviluppatore di una cognizione intelligente, e studiarne dunque i processi guardando alle modalità in cui l'organismo si auto-gestisce tramite scambi con l'esterno e mantenimento dell'organizzazione interna (von Foerster 1960, 1987). Come sottolineato da Damiano (2010: 49), von Foerster ammonisce fin dagli inizi i presupposti e i metodi della neonata scienza cognitiva,

responsabile di aver trasferito lo schema concettuale del calcolatore digitale dal dominio della progettazione ingegneristica di artefatti intelligenti a quello dello studio della cognizione naturale. [...] Sovrapporre il disegno del calcolatore artificiale ai sistemi biologici significa respingere nel retroscena dell'indagine la proprietà specifica che distingue i viventi dai congegni. Vuol dire trattare come un aspetto residuale, essenzialmente trascurabile, la capacità degli organismi di definire e rigenerare da sé la propria organizzazione, senza l'intervento di ingegneri. È incorrere in un accecamento

---

<sup>59</sup> Con entropia si intende la misura del rumore in un sistema. La misura viene adottata da von Foerster seguendo gli studi di Claude Shannon sulla teoria dell'informazione (von Foerster 1960).

<sup>60</sup> Il principio dell'*ordine dal rumore* era stato elaborato da von Foerster (1960) sulla scorta delle riflessioni inaugurate dal fisico Erwin Schrödinger, che aveva formulato il principio dell'*ordine dall'ordine* (Schrödinger 1944) (cfr. Ceruti 1989; Damiano 2010; Capra e Luisi 2016).

teorico decisivo: perdere di vista che la cognizione vivente non consiste in un'attività di *problem-solving* estrinseca, strutturata e imposta da un progettista esterno. Consiste nel produrre soluzioni efficaci nel problema biologico di considerare l'organizzazione in un ambiente mutevole.

Da qui, von Foerster elabora tre fondamentali tesi anti-computazionaliste che condurranno allo sviluppo di una nuova idea di cognizione (Damiano 2010: 50; Ceruti e Damiano 2018: 5):

- 1) L'ambiente non si configura come una riserva di informazioni, ma è invece da considerare come una sorgente di energia e materia, un sistema operazionalmente indipendente ma a cui il sistema biologico deve necessariamente accoppiarsi per alimentare i processi interni;
- 2) Il computer biologico deve essere contemporaneamente aperto a livello termodinamico, in quanto deve scambiare energia con l'esterno, e chiuso a livello informazionale, poiché deve preservare la sua struttura interna;
- 3) L'attività principale del computer biologico non è la computazione, ma la creazione di significato pertinente alla soluzione del problema principale: conservare l'organizzazione.

Gli straordinari contributi di von Foerster non furono gli unici a svilupparsi nel contesto del BCL. Infatti, da lì passano molti dei personaggi che abbiamo già incontrato, come McCulloch, che sembra quasi passare il testimone dalla prima alla seconda cibernetica, o Minsky, che sarebbe restato ammaliato dalle ricerche sulle reti in parallelo costruite allo scopo di simulare il cervello umano, che contribuirono a fargli maturare alcune delle idee e dei ripensamenti che confluirono poi ne *La società della mente* (1985). Inoltre, numerosi furono i talenti che si formarono sotto la guida di von Foerster al BCL: Gordon Pask, Stafford Beer, Humberto Maturana e Francisco Varela sono solamente alcuni dei nomi più celebri nel panorama degli studi della cognizione e dell'artificiale che hanno formato o elaborato le proprie idee in questo contesto e il cui pensiero viene dichiarato più volte esserne debitore.

Considerando questi dati, è interessante notare come, nell'imponente volume *Mind as machine* di Margareth Boden (2006), in più di 1600 pagine, il nome di von Foerster sia citato solamente 8 volte – 4 delle quali in bibliografia, solo una di queste ha von Foerster autore e non curatore: von Foerster (1950-5) – e i riferimenti al BCL siano quasi del tutto assenti. Eppure, come abbiamo brevemente accennato qui e come presente nelle ricostruzioni di numerosi studiosi (Capra e Luisi 2014; Dupuy 1994; Damiano 2010, 2012; Ceruti 1989; Froese 2010; Di Paolo 2003; Luisi 2016), il BCL negli anni '70 è stato il luogo di numerosi incontri eccellenti e la culla di originali idee che sarebbero poi confluite nella storia dell'intelligenza artificiale, come appunto i lavori di Maturana e Varela o le stesse idee sulla cibernetica come epistemologia sviluppate da von Foerster. Sorprende, dunque, che non si sia trovato spazio sufficiente per mettere in luce il contributo di von Foerster in uno dei volumi più ecumenici e di ampio respiro che si è occupato di ricostruire la storia della filosofia

dell'intelligenza artificiale, se non evidenziando gli sforzi poi confluiti nella corrente *mainstream*, come lo studio delle prime reti neurali e la *pattern recognition* (von Foerster 1962; Andrew 2005).

Possiamo provare, alla luce delle conoscenze maturate fino ad oggi, a dare una giustificazione alla mancanza di un approfondimento del contributo di von Foerster nel volume della Boden a partire proprio dalle argomentazioni convocate dall'autrice in merito alla rivoluzione avvenuta nel panorama delle scienze dell'artificiale tra gli anni '70 e '90. Boden stessa, infatti, parla di un movimento di contro-cultura emergente (Boden 2006: 32-36, 852-855) che raccoglie le idee che si oppongono all'impostazione logicista, rappresentazionale e meccanicista delle scienze dell'artificiale ormai classiche, ora affrontata con un impianto che faceva riferimento alla natura biologica e corporea dell'organismo intelligente. Tra questi, un forte accento viene posto sull'opera di costruzione epistemologica di quella che sarebbe poi stata la scuola di Santiago in Cile (Froese 2010), soprattutto nelle personalità di Humberto Maturana e Francisco Varela (Hallowell 2009), che sì, si raccoglie intorno a personalità di origine cilena, ma che si formano nel BCL di von Foerster.

Si può ipotizzare che le radici delle motivazioni di tale 'dimenticanza' siano rintracciabili in almeno due ambiti distinti: la prima è una motivazione politica, per cui le idee che volevano evidenziare la natura biologica della cognizione venivano relegate a movimento di controtendenza rispetto alla corrente maggioritaria, il cui impianto teorico sembrava essere ormai troppo lontano anche per essere solamente messo a confronto con questa e le discussioni specialistiche vertevano sui metodi di miglioramento, non di sostituzione, delle tecnologie e metodologie impiegate (cfr. Boden 2006: 858-861 sul contributo di Drew McDermott alle critiche interne all'AI); l'altra motivazione, comunque collegata alla prima, si svela nella (im)maturità delle tecnologie del tempo, le quali erano sviluppate seguendo un'altra impostazione filosofica e cognitiva e l'implementazione di principi provenienti dalla biologia organica si mostrava come una chimera difficilmente afferrabile tramite quei mezzi. In sostanza, senza voler imputare a Margareth Boden di aver omesso i crediti dovuti a von Foerster, ciò che salta all'occhio è una narcotizzazione dei tratti teorici originari delle idee afferenti alla contro-cultura dell'intelligenza artificiale, a favore d'altro canto di un focus mirato sullo sviluppo delle tecnologie e un necessario approfondimento di quelle idee che ne hanno permesso la costruzione.

Uno dei motivi per cui Boden introduce il discorso sulla svolta biologica nelle scienze cognitive e in quelle dell'artificiale è l'adesione al movimento contro-culturale di un personaggio la cui opinione in merito non poteva essere ignorata. Infatti, Terry Winograd era già celebre nell'ambito della GOF AI per aver creato SHRDLU (§1.3.1.3, Winograd 1972), aveva inaugurato il convegno IJCAI del 1971, ed era considerato ampiamente come uno degli esponenti di spicco della GOF AI simbolica. Abbiamo visto già come SHRDLU sia stato importante in quanto una delle prime attestazioni della possibilità

di funzionare delle macchine programmate seguendo la teoria simbolica computazionale: anche se per Winograd il suo programma non si configurava come un esperimento sintetico sulle funzioni cognitive legate alla comprensione linguistica nell'essere umano, questo non scoraggiò gli psicologi cognitivi dal considerare SHRDLU come una prova delle teorie impiegate per costruirlo. In sostanza, Winograd era percepito come uno dei pionieri del NLP, ma ciò non lo fermò dal cambiare completamente idea e dall'avanzare una critica all'AI simbolica che, forse proprio per questo, fece trasalire gli esperti della GOFAI sulle loro poltrone, i quali considerarono quello di Winograd un vero e proprio tradimento<sup>61</sup>. Quello che fece davvero la differenza per Winograd furono una serie di incontri avvenuti nel corso degli anni '70, che gli fecero riconsiderare il concetto di cognizione: il primo, avvenuto con le idee elaborate da Dreyfus riguardo la fenomenologia, lo portò a pensare la cognizione come un fenomeno impossibile da indagare senza prendere in considerazione l'esperienza incarnata nel mondo; l'altro, avvenuto con l'ingegnere ed economista cileno Fernando Flores, lo introdusse agli studi cibernetici sulla cognizione del connazionale Humberto Maturana, che condussero Winograd a riconsiderare la cognizione come un fenomeno fondamentalmente biologico.

Parleremo dell'argomentazione di Dreyfus nel paragrafo successivo (cfr. §2.3.1.3), anche sottolineando i caratteri comuni con le idee di Winograd. Per ora, ci interessa indagare come sia proprio attraverso l'opera di Winograd e Flores che una delle teorie oggi più diffuse riguardo la natura biologica della cognizione abbia acquistato validità nel dominio delle scienze cognitive e dell'artificiale ad ampio raggio, arrivando a informare gli approcci alternativi della robotica negli anni successivi.

Nel volume *Understanding Computers and Cognition* (1986), Winograd e Flores dichiarano fin dalle prime pagine il loro malcontento nei confronti dell'approccio simbolico computazionalista e aprono una nuova linea di indagine basata su presupposti lontanissimi:

All new technologies develop within the background of a tacit understanding of human nature and human work. The use of technology in turn leads to fundamental changes in what we do, and ultimately in what it is to be human. We encounter the deep questions of design when we recognize that in designing tools we are designing ways of being. (Winograd e Flores 1986: xi)

[...] We are not concerned with providing new answers to questions about technology as they have traditionally been posed. We look towards new questions that can lead to the design and use of machines that are suited to human processes. (Ibid: xiii)

Much of our book is an attempt to show the non-obviousness of the rationalistic orientation and to reveal the blindness that it generates. (Ibid.: 17).

---

<sup>61</sup> Il "tradimento" di Winograd è stato paragonato all'altrettanto scandaloso voltafaccia operato da Putnam (Hacking 1989) qualche anno prima nei confronti del funzionalismo, con l'aggravante che quest'ultimo non solo aveva contribuito a svilupparlo, ma era stato addirittura il primo a formalizzarlo (Putnam 1960).

Seguendo le intuizioni di Weizenbaum, anche Winograd e Flores puntano subito l'attenzione sul fatto che l'adozione non problematizzata della metafora computazionale possa condurre molto lontano da certe importanti questioni. Il problema principale sembra, secondo gli autori, essere il fatto che tale concettualizzazione non supporta una comprensione adeguata di cosa i dispositivi artificiali come i computer facciano effettivamente nel contesto della pratica umana. Per la prima volta, così, vediamo introdotta nel dibattito sull'intelligenza artificiale la questione pragmatica della pratica di utilizzo, declinata nello specifico al design delle nuove tecnologie informate da un'analisi interdisciplinare sulle modalità di impiego di questi strumenti nella pragmatica umana dell'azione.

La proposta innovativa di Winograd e Flores per ripensare il campo dell'intelligenza artificiale consiste, in primo luogo, in «reveal[ing] the implicit understanding of human language, thought, and work that serves as a background for developments in computer technology» (Winograd e Flores 1986: 7). L'operazione di decostruzione della metafora computazionalista ha dunque inizio, sostituendo innanzitutto i presupposti abbracciati dalla GOFAI, come il dualismo mente-corpo, la struttura logica computazionale della cognizione e l'intelligenza come attività gerarchica di *problem-solving*<sup>62</sup>, con nuove idee, che potessero indirizzare l'indagine in modi fino ad allora intentati.

In particolare, la teoria adottata nel capitolo "Cognition as a biological phenomenon" (Ibid: 38-53) è quella dell'*autopoiesi*, concepita da Maturana (1970) e poi elaborata con Francisco Varela e applicata alla cognizione (Maturana e Varela 1972, 1980). Al tempo, Maturana era già famoso nel campo della biologia teorica per aver sviluppato, insieme ad alcuni dei suoi grandi maestri, un approccio cibernetico alla visione mettendolo in relazione alla neurofisiologia<sup>63</sup> (Lettvin et al. 1959), ma era considerato oscuro nella formulazione delle idee e di difficile comprensione per i non addetti ai lavori<sup>64</sup>. Tuttavia, Winograd e Flores compresero a pieno la nuova concettualizzazione autopoietica, forse anche grazie alle mediazioni delle teorie cibernetiche di Stafford Beer<sup>65</sup> a cui Flores era interessato, e la integrarono nel loro approccio alternativo alla progettazione dell'intelligenza artificiale. Ciò che interessava maggiormente ai due autori era vedere come la riconcettualizzazione del dominio cognitivo nei termini in cui era proposta da Maturana e Varela

---

<sup>62</sup> Durante la sua carriera di scienziato cognitivo e informatico nella GOFAI, Winograd aveva apportato notevoli cambiamenti alla programmazione, come ad esempio il concetto di programmazione processuale in sostituzione (o connubio) a quella dichiarativa, spesso accostate alle categorie di *know that* e *know how* elaborate da filosofo Gilbert Ryle, e alla concezione gerarchica del programma (concependo e implementando la struttura eterarchica) (Boden 2006: 777-778).

<sup>63</sup> Il concetto centrale dell'articolo, per come descritto dagli autori: «[...] the eye speaks to the brain in a language already highly organized and interpreted, instead of transmitting some more or less accurate copy of the distribution of light on the receptors» (Lettvin et al. 1959: 251).

<sup>64</sup> Nelle parole di Margareth Boden (2006: 855): «Maturana was rarely even read, partly because he wrote in such an opaque manner; his work became popular only much later [...]».

<sup>65</sup> Stafford Beer, oltre ad essere stato un esponente di spicco della seconda cibernetica e attivo nel BCL fondato da von Foerster, è l'autore della prefazione a *Autopoiesi. L'organizzazione del vivente* (Maturana e Varela 1973), poi confluita in *Autopoiesi e cognizione. La realizzazione del vivente* (Maturana e Varela 1980).

potesse essere applicata alla progettazione di computer e robot. Per capire bene, dunque, come è avvenuta questa operazione, prendiamo in considerazione in breve gli argomenti sviluppati da Maturana e Varela per definire la teoria autopoietica.

La teoria dell'autopoiesi si posiziona in aperto contrasto con le teorie computazionaliste, non olistiche e logiciste riguardo la vita e la mente. Nella sua definizione più chiara, l'autopoiesi è

A network of processes of production (transformation and destruction) of components that produces the components that: (i) through their interactions and transformations continuously regenerate the network of processed (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its relation as such a network. (Maturana e Varela 1980: 79)

In sostanza, ciò che fa il sistema autopoietico è mantenere costante la sua organizzazione e definire i propri confini attraverso una continua produzione delle sue componenti. Se l'autopoiesi si interrompe, l'organizzazione del sistema – la sua identità come un particolare tipo di unità – è persa e il sistema si disintegra. Se vista nei termini di una descrizione puramente teorica, l'autopoiesi può essere identificata in qualsiasi macchina che si auto-determina e acquista una propria autonomia<sup>66</sup> nel processo di auto-costituzione in relazione all'ambiente. In pratica, l'autopoiesi è l'esatta descrizione di una struttura tipica di un solo tipo di sistema: quello vivente.

I sistemi viventi sono i sistemi cognitivi, e il vivere in quanto processo è un processo di cognizione. (Maturana 1970: 59)

La caratteristica essenziale dei sistemi viventi, gli unici che hanno sviluppato forme di cognizione intelligente, è allora quella di essere una collezione di componenti che costituiscono un'unità che può vivere o morire. L'osservazione centrale di Maturana è che esattamente questa semplice proprietà conduce ai fenomeni complessi della vita: una macchina autopoietica che esiste in uno spazio fisico è un sistema vivente. È interessante, allora, vedere come il meccanicismo in questa sede si trasformi ancora, portandoci a rileggere sotto la lente autopoietica il concetto di macchina che abbiamo affrontato in §1.2, come fatto da Luisa Damiano (Damiano 2010: 146)

Qui va collocata e letta la dichiarazione di adesione allo stile esplicativo del meccanicismo fatta dalla Scuola di Santiago. [...] è una forma di esplicazione che riprende l'idea di macchina, ma non la elabora secondo i canoni del meccanicismo classico. La ricostruisce con nozioni olistiche come quella di organizzazione, ma senza scivolare nel vitalismo. Si situa in una dimensione conoscitiva in cui il lavoro scientifico supera l'alternativa tra il prevedere esaustivamente e l'intuire senza poter esplicitare. In

---

<sup>66</sup> Il concetto di autonomia è quello da cui Maturana e Varela fanno iniziare la loro indagine sulla struttura della cognizione e della vita: «L'autonomia quale caratteristica essenziale del vivente è un fatto talmente immediato che quando osserviamo qualcosa che ci appare autonomo la nostra reazione spontanea è di considerarlo vivo» (Maturana e Varela 1973: 23). Come sottolineato da Damiano (2010: 166), l'autonomia è una caratteristica fondamentale del processo autopoietico, descrivibile come una «*triade concettuale* che iscrive l'una nell'altra le proprietà biologiche dell'auto-stabilizzazione [simultanea chiusura e apertura del sistema], l'auto-produzione [dei propri componenti, autopoiesi] e l'auto-separazione da uno sfondo [identità]».



quest'area *conoscere scientificamente* - «spiegare» - diventa *fare – costruire, creare, fabbricare*.

A differenza della teoria computazionalista, quindi, in cui la cognizione si configura come un processo discreto, composto da moduli e funzioni che non comunicano direttamente tra di loro, in cui la percezione è atta a raccogliere input dall'esterno per sottoporli al processo cognitivo computazionale che ne elabora degli output, la cognizione nella teoria autopoietica è l'attività stessa che il sistema realizza nel suo costante tentativo di mantenere stabile la propria organizzazione (Maturana e Varela 1980). Ciò è impossibile senza scambi con l'esterno, come imparato già dalla lezione di von Foerster, secondo cui l'ambiente non è considerato come una fonte di informazioni ma come un elemento necessario, un *medium* imprescindibile nell'equazione della cognizione.

Attraverso questa breve introduzione alla teoria dell'autopoiesi, possiamo ora apprezzare lo sforzo interpretativo compiuto da Winograd e Flores per estrapolare le caratteristiche di un nuovo approccio alle scienze dell'artificiale. In particolare, Winograd e Flores (1986) argomentano a favore di almeno tre riconfigurazioni del dominio della progettazione degli artefatti artificiali, rese possibili dalle teorie di Maturana e Varela:

- i) La teoria dell'autopoiesi consente di *reformulare il rapporto tra percezione e azione*. Infatti, invece di concepire la percezione come attività completamente indipendente o vicaria della cognizione e in quanto processo passivo di raccolta di input informativi dall'esterno, la percezione diventa un processo proattivo strettamente legato all'azione e al mantenimento dell'organizzazione interna dell'organismo. È importante in questo contesto la nozione di *accoppiamento strutturale*, ovvero il fatto che le caratteristiche dell'organismo, sia corporee che funzionali, sono selezionate dalle caratteristiche dell'ambiente, cioè quelle caratteristiche che l'organismo individua come funzionali al mantenimento della sua organizzazione interna. Ma non solo: l'accoppiamento strutturale implica anche che siano le caratteristiche dell'organismo ad essere selezionate dalle caratteristiche dell'ambiente, istituendo così un rapporto di co-dipendenza strettissimo che definisce il sistema organismo-ambiente come unità osservabile. La percezione, quindi, non si mostra più come quella capacità di afferrare una realtà esterna all'organismo, a lui preesistente e indipendente in termini assoluti; significa, invece, *costruire* una realtà in cui l'organismo è capace di agire date le sue caratteristiche strutturali e organizzazionali.
- ii) *Comporta il rifiuto della componente rappresentazionale della spiegazione cognitiva*: l'elaborazione del concetto di accoppiamento strutturale tra l'organismo e il suo ambiente è ciò che permette di eliminare la funzione rappresentazionale della cognizione. Winograd e Flores parlano di quella che Maturana chiama *fallacy of instructive interaction*, attraverso cui viene descritta la credenza diffusa che nelle interazioni con l'ambiente l'organismo ne

acquisisca una rappresentazione, estrapolandone informazioni che vengono poi mappate nella struttura del sistema nervoso. Tuttavia, le idee di Maturana e Varela permettono di riconfigurare il rapporto guardando ai cambiamenti strutturali come pattern di attività in risposta a una perturbazione, e non come rappresentazioni del mondo dell'agente. Le rappresentazioni allora diventano superflue nel momento in cui la sola attività fondamentale portata avanti dall'organismo è quella di agire efficacemente all'interno del suo mondo, districandosi in una catena interazionale di perturbazioni e tentativi di stabilizzazione, senza doversi creare alcun modello su cui ragionare in maniera astratta. Di conseguenza, anche processi cognitivi considerati fino a quel momento strettamente dipendenti dal modello rappresentazionale, come l'apprendimento o la memoria, possono essere inquadrati in una nuova formulazione:

Learning is not a process of accumulation of representations of the environment; it is a continuous process of transformation of behavior through continuous change in the capacity of the nervous system to synthesize it. Recall does not depend on the indefinite retention of a structural invariant that represents an entity (an idea, image or symbol), but on the functional ability of the system to create, when certain recurrent conditions are given, a behavior that satisfies the recurrent demands or that the observer would class as reenacting of a previous one. (Maturana 1970: 45)

iii) *La cognizione è vista come un processo in un dominio biologico:*

La cognizione è un fenomeno biologico e lo si può capire solo come tale: qualsiasi indagine epistemologica nel dominio della conoscenza richiede questa comprensione. (Maturana 1970: 51)

Per Winograd e Flores, ciò comporta la consapevolezza riguardo a diversi modi di porre il discorso sulla possibilità delle scienze dell'artificiale per come sviluppata fino a quel momento di creare intelligenza artificiale, almeno nel senso forte del termine, ma permette anche di aprire nuove concettualizzazioni pratiche della relazione fra umani e tecnologie. Questo comporta una riflessione profonda sul ruolo del designer di artefatti artificiali e del programmatore informatico, che devono essere consapevoli delle potenzialità della tecnologia, delle conoscenze precedenti che informano la costruzione di programmi per computer e delle aspettative con cui le persone si avvicinano alle tecnologie che loro stessi costruiscono:

In Maturana's terms, the key to cognition is the plasticity of the cognitive system, giving it the power of structural coupling. As the domain of interaction is modified, the structure of the interacting system changes in accord with it. We cannot directly impose a new structure on any individual, but whether we design changes to the space of interactions, we trigger changes in individual structure – changes to the horizon that is precondition for understanding. (Winograd e Flores 1986: 178).

L'operazione di sostituzione della metafora computazionalista con il modello autopoietico della cognizione, allora, comporta non solo un ripensamento riguardo le ipotesi cognitive fino a quel momento utilizzate per costruire artefatti artificiali, come quelle relative alla percezione o alle modalità di cognizione, ma anche la necessità di prendere in considerazione il ruolo dell'osservatore, del designer, per comprendere pienamente come poter sfruttare i meccanismi cognitivi per far funzionare adeguatamente le tecnologie artificiali. La riflessione sulla costruzione di un dominio interno, nel robot<sup>67</sup> ad esempio, viene affrontata prendendo in considerazione i principi della teoria autopoietica e applicandoli al rapporto che c'è tra designer e il suo prodotto informatico:

The most successful designs are not those that try to fully model the domain in which they operate, but those that are 'in alignment' with the fundamental structure of that domain, and that allow for modification and evolution to generate new structural coupling. As observers (and programmers), we want to understand to the best of our ability just what the relevant domain of action is. This understanding guides our design and selection of structural changes, but need not (and in fact cannot) be embodied in the form of the mechanism. (Winograd e Flores 1986: 53).

Da qui, si concretizza la critica alla concettualizzazione del pensiero come manipolazione di rappresentazioni in un dominio astratto costruito appositamente come un monito per i programmatori, i quali devono essere consapevoli delle proprie azioni e guardare al modo in cui la loro attività può rinforzare o indebolire la visione della mente che prendono in riferimento.

Our criticism of descriptions of human thought as 'decision making' and language understanding as the manipulation of representations is not just a prediction that certain kinds of computer program will fail. It reflects a deeper concern with the discourse and actions that are generated by a rationalistic interpretation of human action. Computer systems can easily reinforce this interpretation, and working with them can reinforce pattern of acting that are consistent with it. On the other hand, where there is a danger there is an opportunity. We can create computer systems whose use leads to a better domain of interpretation. The machine can convey a kind of 'coaching' in which new possibilities for interpretation and action emerge. (Winograd e Flores 1986: 178-179).

Lo sguardo rivolto alla pratica di sviluppo e utilizzo delle tecnologie artificiali è indirizzata dalla teoria autopoietica verso la comprensione del processo di interazione con il mondo, il quale include anche gli oggetti con cui interagiamo. Ma la questione della comprensione della pratica d'uso viene considerata sotto un aspetto nuovo, ricalcato dalla filosofia fenomenologica heideggeriana, cioè quello di *breakdown*.

For Heidegger, 'things' emerge in breakdown, when unreadiness-to-hand unconceals them as a matter of concern. Maturana sees the presence of objects and properties as relevant only in a domain of distinctions made by an observer. In the domain of biological mechanism, they do not exist. Both authors recognize that we are situated in a world that is not of our own making. Their central insight is that this world, constituted as a world of

---

<sup>67</sup> Un esplicito giudizio in riferimento alla robotica della GOFAI è espresso da Winograd e Flores in relazione alla concettualizzazione di percezione e azione: «Perception is much more than the encoding of an external world into a representation, and action is more than the execution of 'motor routines'». (Winograd e Flores 1986: 128)

objects and properties, arises only in the concernful activities of the person. (Winograd e Flores 1986: 73)

A breakdown is not a negative situation to be avoided, but a situation of non-obviousness, in which the recognition that something is missing leads to unconcealing [...] some aspects of the network of tools that we are engaged in using. A breakdown reveals the nexus of relations necessary for us to accomplish our task. (Winograd e Flores 1986: 165).

Nel processo di utilizzo del computer o del programma, accadono necessariamente delle rotture derivanti dalla non corrispondenza tra l'intento e la pratica d'utilizzo: sono questi i momenti che devono essere osservati, poiché svelano caratteristiche attese e modalità di approccio alla tecnologia da parte dell'individuo e instradano l'interpretazione tramite cui l'artefatto deve essere progettato per essere adeguato al task guardando principalmente all'azione<sup>68</sup> (cfr. §5.1.1). Winograd e Flores, tenendo insieme la teoria dell'autopoiesi di Maturana e la fenomenologia heideggeriana, costruiscono quindi un approccio in cui lo sguardo è rivolto alle modalità di mettere in evidenza la pragmatica d'uso dell'artefatto digitale, considerando una serie di osservazioni riguardo alla natura della mente e al dispiegamento della pratica di utilizzo che non erano fino ad allora entrate nel dibattito sull'intelligenza artificiale.

Componendo un approccio attraverso l'utilizzo di fonti quasi mai prese in considerazione nel dibattito sull'intelligenza artificiale e di difficile comprensione ai non specialisti, il discorso risulta altamente originale, tanto che influenzerà profondamente alcune figure che saranno centrali nello sviluppo di una robotica alternativa a quella informata dalla GOFAI: primo fra tutti Randall Beer, fautore, insieme a Rodney Brooks, di un approccio alternativo alla robotica basato sulle idee di cognizione propugnate in questo libro (Beer 2004). C'è da notare, tuttavia, che alcune istanze erano più originali di altre: infatti, se la teoria dell'autopoiesi non aveva ancora mai fatto la sua apparizione nel dibattito 'di maggioranza' sulla cognizione artificiale, le idee di Heidegger e le potenzialità che la sua riflessione poteva far emergere se importata nel discorso sull'intelligenza artificiale erano state introdotte alla comunità di *computer scientist* e programmatori più di dieci anni prima da Hubert Dreyfus. Winograd avrebbe condiviso molti pranzi con Dreyfus negli anni '70 e, evidentemente, le conversazioni che ebbero influenzarono profondamente la riflessione di Winograd riguardo una necessaria riconcettualizzazione della questione dell'intelligenza e della cognizione che includesse la natura esperienziale e corporea della cognizione umana.

---

<sup>68</sup> «When we look at computer technology this basic point guides our understanding in several ways. First, it is critical in our anticipation of the kinds of computer tools that will be useful. In a tradition that emphasizes thought as an independent activity, we will tend to design systems to work within that domain. In fact much of the current advertising rhetoric about computers stresses the role they will play in 'applying knowledge' and 'making decisions'. If, on the other hand, we take action as primary, we will ask how computers can play a role in the kinds of actions that make up our lives – particularly the communicative acts that create requests and commitments and that serve to link us to others.» (Winograd e Flores 1986: 71-72)

Secondo Winograd, le conseguenze che l'importazione della teoria dell'autopoiesi avrebbero avuto sul dominio dell'intelligenza artificiale dovevano andare di pari passo con le riconsiderazioni che Dreyfus aveva iniziato a suggerire già da qualche anno e che aveva contribuito a creare quella contro-cultura di cui sia Weizenbaum, sia Winograd e Flores facevano parte.

### 2.3.1.3 Il corpo e l'esperienza: Hubert Dreyfus

A dare l'inizio alla corrente che critica l'impostazione teorica e pratica della GOFAI è stato il filosofo Hubert Dreyfus. Dreyfus è diventato celebre come critico della GOFAI a seguito della pubblicazione del suo libro *What computers can't do* (1972), in cui esponeva, senza limitare i tecnicismi, le ragioni per cui il progetto dell'intelligenza artificiale per come era stato attuato fino a quel momento era secondo lui destinato al fallimento. Il libro diventò fin da subito un *best seller* e si può imputare a lui (almeno in una parte sostanziale) il fatto che il dibattito sull'intelligenza artificiale, in quel periodo, abbandonò l'esclusività delle aule di informatica e si insinuò anche in quelle di filosofia delle maggiori università americane, come il MIT e Stanford. Il modo in cui Dreyfus aveva posto la questione era accessibile non più solamente agli esperti programmatori, ma dimostrava avere un interesse anche per gli studiosi di discipline umanistiche, tanto da raccogliere intorno a sé una gran parte di quelli che si dimostravano scettici in merito al futuro dell'AI in entrambi i campi.

Sono celebri le predizioni fatte da Dreyfus nel suo libro, le quali si caratterizzano per un forte pessimismo nei confronti di qualsiasi realizzazione degli obiettivi posti dai ricercatori della corrente simbolico-computazionalista, come nel caso dei programmi di scacchi o dei primi sistemi esperti (Dreyfus 1965: 6). La maggior parte degli esponenti della corrente *mainstream* dell'AI e delle scienze cognitive computazionaliste, contrari alle idee di Dreyfus, si concentreranno sulle predizioni che sarebbero state smentite di lì a poco, come quella riferita all'impossibilità di creare un programma giocatore di scacchi che fosse all'altezza anche di un dilettante, quando Dreyfus stesso fu sconfitto l'anno dopo (1966) proprio da uno di quei programmi (Boden 2006: 840). Tuttavia, alcune delle critiche formulate da Dreyfus a cavallo tra la fine degli anni '60 e l'inizio degli anni '70 risultano valide ancora oggi, e sicuramente sono quelle che hanno contribuito principalmente alla creazione di una corrente alternativa a quella della GOFAI sia per i programmi di intelligenza artificiale, ma ancora di più per la riconfigurazione della robotica e delle scienze cognitive.

Le idee di Dreyfus riguardo all'inadeguatezza dei metodi utilizzati fino a quel momento per costruire artefatti artificiali che potessero somigliare o riprodurre funzioni cognitive umane si basano su un'osservazione che lega l'impresa dell'IA alle riflessioni filosofiche precedenti. Infatti, Dreyfus nota che gli studi in quel momento appartenenti alla corrente maggioritaria dell'intelligenza artificiale

potevano essere inquadrati come l'ultimo esempio di una corrente di pensiero che attraversa diversi secoli, appartenente ad un dominio filosofico preciso:

[...] the pioneers in CS [Cognitive Simulation] had learned a lot, directly and indirectly, from the philosophers. They had taken over Hobbes's claim that reasoning was calculating, Descartes' mental representations, Leibniz's idea of a "universal characteristic" – a set of primitives in which all knowledge could be expressed – Kant's claim that concepts were rules, Frege's formalization of such rules, and Russell's postulation of logical atoms as the building blocks of reality. In short, without realizing it, AI researchers were hard at work turning rationalist philosophy into a research program. (Dreyfus 2008: 332-333)

La ricerca nel campo dell'intelligenza artificiale, dunque, era l'ultimo (in termini temporali) risultato della corrente razionalista in filosofia, intenta a spiegare il mondo in atomi e relazioni tra questi, con la fiducia nelle capacità razionali di pensiero tipica di questa corrente. Tuttavia, Dreyfus fa un passo ulteriore e costruisce un parallelismo tra la storia della filosofia e quella che avrebbe atteso i teorici dell'intelligenza artificiale:

[...] I began to suspect that the critical insights formulated in the existential armchairs, especially Heidegger's and Merleau-Ponty's, were bad news for those working in AI laboratories – that, by combining rationalism, representationalism, conceptualism, formalism, and logical atomism into a research program, AI researchers had condemned their enterprise to reenact a failure. (Ibid.: 333)

L'intera impresa dell'intelligenza artificiale era condannata al fallimento: come avevano mostrato i filosofi 'continentali' citati, appartenenti in gran maggioranza alle correnti fenomenologiche ed esistenzialiste, nella critica ai loro stessi predecessori appartenenti alla corrente razionalista e positivista, ci sono numerosi altri fattori da tenere in considerazione quando si cerca di trovare una soluzione al problema della cognizione e del linguaggio, primi fra tutti il corpo e l'esperienza. Come spesso affermato nei suoi scritti, secondo Dreyfus i progetti dell'intelligenza artificiale si erano dimostrati come le più efficaci prove dell'inadeguatezza della teoria che implementavano:

We can then view recent work in artificial intelligence as a crucial experiment disconfirming the traditional assumption that human reason can be analyzed into rule-governed operations on situation-free discrete elements – the most important disconfirmation of this metaphysical demand that has ever been produced. This technique of turning our philosophical assumptions into technology until they reveal their limits suggests fascinating new areas for basic research. (Dreyfus 1972: 216)

Il nucleo centrale di queste idee può essere rintracciato già a partire dalla metà degli anni '60 nel suo testo *Alchemy and Artificial Intelligence* (Dreyfus 1965). Il documento, scritto come una relazione dopo un viaggio fatto presso il RAND, istituto di ricerca dell'università di Stanford in cui lavorava Newell e a cui Simon faceva spesso visita, era una prima attestazione delle difficoltà affrontate dalle tecnologie artificiali a lui contemporanee nello svolgere i compiti per cui erano stati creati. In particolare, nel documento del '65 troviamo già delineate alcune delle osservazioni che

verranno poi riprese e sviluppate nel libro del '72: da una parte, l'intero progetto dell'intelligenza artificiale risultava di impossibile attuazione per via di ragioni filosofiche – ovvero perché le forme più alte di intelligenza sono derivate da forme più basse che riguardano l'azione corporea; dall'altra, la critica alle tecnologie artificiali dell'epoca si sviluppa attraverso un'analisi serrata dei presupposti filosofici e cognitivi, anche nascosti, che ne hanno informato l'implementazione. Le quattro accuse formulate qui da Dreyfus (*realiance on the fringe of consciuosness, discrimination between the essential and the accidental, tolerance of ambiguity, perspicuous grouping*) riguardavano ognuna un aspetto fondamentale dell'intelligenza umana che la GOFAI aveva escluso dalle proprie considerazioni e costituiscono la prima impalcatura teorica su cui la corrente contro-culturale, a cui abbiamo accennato prima, costruisce la propria invettiva. In sostanza, l'accusa generale è riassumibile nel titolo del documento: l'intelligenza artificiale era una moderna riattualizzazione della pratica alchemica, che poteva condurre fortuitamente a qualche risultato degno di nota, ma che in realtà si basa su presupposti errati e incompleti e porta con sé anche una cospicua dose di pensiero magico. Il fatto che l'AI si sia concentrata sulla replicazione dell'intelligenza formale “semplice” e “associazionista” non può nascondere il fatto che altri tipi di intelligenza, come quella formale “complessa” e quella “non-formale”, non possano essere formalizzate secondo i dettami della GOFAI, dal momento che le caratteristiche essenziali delle ultime due forme, come l'intuito alla base dell'intelligenza non-formale o l'esperienza, sono strettamente legate alla corporeità dell'agente umano inserito nel suo mondo:

It seems self-evident that we could simulate intelligent behavior if we could build or simulate a device which functioned exactly like the human brain. But even this could be challenged if it could be shown that the body plays a crucial role in making possible intelligent behavior. (Dreyfus 1965: 59)

Nel testo del 1972, l'argomentazione si potenzia e Dreyfus sviluppa una approfondita giustificazione filosofica per gli attacchi che erano già stati sferrati nel documento del '65. In particolare, Dreyfus articola un'analisi in quattro assunti:

- *Assunto biologico*. Dreyfus riassume in questi termini quello che fino ad ora abbiamo trattato come metafora computazionale, definendo tale idea come «the naïve assumption that man is a walking example of a succesful digital computer program» (Dreyfus 1972: 71). Citando Lettvin, Dreyfus puntualizza che le ricerche sulla natura fortemente interazionale della struttura cerebrale suggeriscono che la possibilità di produrre intelligenza tramite computer digitali è molto lontana, se non irrealizzabile (Ibid.: 74).
- *Assunto Psicologico*. In questo assunto Dreyfus sintetizza il presupposto che la cognizione funzioni come un processo di manipolazione di simboli, in cui le informazioni vengono trasformate attraverso processi logici e computazionali. Per Dreyfus, tale assunto non tiene

conto della complessità e della natura situazionale della cognizione umana, influenzata da diversi fattori come le emozioni, l'esperienza corporea e l'interazione sociale, considerati fattori non simbolici. Oltretutto, viene messo in discussione il nucleo fondamentale del concetto di mente come "processing information": «Although psychologists describe that function called the mind as "processing information", this does not mean that it actually processes information in the modern technical sense, nor that it functions like a digital computer, that is, that it has a program» (Ibid.: 77).

- *Assunto Epistemologico*. Viene descritto come la posizione secondo cui

although human *performance* might not be *explainable* by supposing that people are actually following heuristic rules in a sequence of unconscious operations, intelligent *behavior* may still be *formalizable* in terms of such rules and thus reproduced by a machine. (Ibid.: 101)

Nella visione data da Dreyfus, questo assunto è necessario per pensare alla cognizione come un processo di ricerca nello spazio del problema, ovvero una sorta di motore di ricerca che elabora i dati attraverso algoritmi per arrivare a conclusioni razionali. Attraverso una disamina mirata alle posizioni di Minsky riguardo la possibilità di formalizzazione di aspetti non formali dell'intelligenza umana, che secondo Dreyfus ricalcano perfettamente l'assunto epistemologico (Ibid.: 103-109), critica la posizione astratta assunta dai ricercatori della GOFAI mostrando come l'intelligenza umana sia un'intelligenza pratica, molto abile a districarsi nell'ambiguità (cosa in cui l'AI pecca puntualmente), e ancorata a un sostrato esperienziale e corporeo non eliminabile se si vuole arrivare a comprenderla.

- *Assunto Ontologico*. Definisce la credenza che «the world can be exhaustively analyzed in terms of determinate data or atomic facts» (Ibid.: 117). Ciò implica che esista una realtà oggettiva, potenzialmente catalogabile, e che l'appropriata forma di descrizione possa aprire la possibilità di realizzare artefatti intelligenti con capacità umane. Per sottolineare l'inadeguatezza di questa visione del mondo, Dreyfus si appoggia alle critiche già sviluppate in filosofia da Wittgenstein, Heidegger e Merleau-Ponty, i quali avevano costruito un apparato critico della questione ormai classica nel loro dominio di riferimento (Ibid.: 124-125).

A partire da questa analisi, Dreyfus elabora alcune alternative con cui sostituire gli assunti della GOFAI dimostrati essere inadeguati per fondare una scienza dell'artificiale, motivandola attraverso un cambio di paradigma filosofico:

It is not some specific explanation, then, that has failed, but the whole conceptual framework which assumes that an explanation of human behavior can and must take the Platonic form, successful in physical explanation; that situations can be treated like physical states; that the human world can be treated like the physical universe. If this



whole approach has failed, then in proposing an alternative account we shall have to propose a different *sort* of explanation, a different sort of answer to the question “How does man produce intelligent behavior?” or even a different sort of questions, for the notion of “producing” behavior instead of simply exhibiting it is already colored by the tradition. (Ibid: 145)

Dreyfus, dunque, fa riferimento al paradigma filosofico già citato in precedenza, composto principalmente dalle posizioni di Wittgenstein, della psicologia della Gestalt, di Merleau-Ponty e di Heidegger, oltre che al contributo del filosofo della scienza Polanyi, attraverso le riflessioni dei quali costruisce un approccio in cui l’intelligenza umana è ancorata a e radicata nel corpo, il corpo è ancorato nel mondo, e insieme il corpo e il mondo producono l’esperienza, che permette all’essere umano di sviluppare tutte le forme di intelligenza, formale e non formale, che dimostra nel suo comportamento. In sostanza, le posizioni di Dreyfus si focalizzano su:

- *Il ruolo del corpo nel comportamento intelligente*: a partire dalla critica al dualismo cartesiano tra materialità del mondo e immaterialità dell’anima, Dreyfus attribuisce al corpo una funzione fondamentale, imprescindibile, per la possibilità dell’intelligenza umana:

A brain in a bottle or a digital computer might still not be able to respond to new sorts of situations because our ability to be in a situation might depend, not just on the flexibility of our nervous system, but rather on our ability to engage in practical activity. After some attempts to program such a machine, it might become apparent that what distinguishes persons from machines, no matter how cleverly constructed, is not a detached, universal, immaterial soul but an involved, self-moving, material body. (Ibid.: 148).

Seguendo questa argomentazione, alcune capacità mostrate dall’essere umano non potranno mai essere realizzate su un dispositivo artificiale, data la specificità del sostrato corporeo dell’esperienza umana: è il caso, per Dreyfus, della *pattern recognition* che, per quanto possa dare buoni risultati se formalizzata, non si riuscirà ad eguagliare la capacità umana, poiché strettamente collegata al modo umano di essere situati nel mondo.

- *L’intelligenza umana si dà sempre in una situazione*: riprendendo la nozione di *orizzonte* sviluppata da Heidegger, Dreyfus sottolinea che mondo di un essere umano è sempre pre-strutturato in termini di scopi e rilevanze che rendono alcuni oggetti e percezioni significative in quanto tali, dal momento che sono significative per l’individuo. Ciò è impossibile per un computer o un robot, dal momento che questi possono avere a che fare solamente con oggetti determinati dal programmatore, il quale può solamente provare a simulare il campo di significati che possono aprirsi nell’interazione con il mondo stesso (Ibid.: 174)

Soprattutto riguardo questo ultimo punto, è interessante notare come l’impostazione chiamata in gioco da Dreyfus abbia fornito un punto di vista alternativo riguardo uno dei problemi che erano stati formulati a partire dall’impostazione della GOFAI: il *frame problem*.

Il *frame problem* è un problema classico dell'intelligenza artificiale, formalizzato in prima istanza da McCarthy e Hayes (1969). Nonostante la formalizzazione originale avvenga in ambiti diversi da quello epistemologico (Shanahan 2004), il problema del *frame* acquista molto velocemente una rilevanza nell'ambito della creazione di un'intelligenza artificiale che non dovesse solamente avere a che fare con mondi e problemi giocattolo, ma che fosse in grado di districarsi nella complessa trama della realtà materiale. Il primo a importare il discorso nell'ambito epistemologico delle scienze cognitive fu Daniel Dennett (1978), formulandolo come segue:

Quando una creatura cognitiva, un'entità che ha molte credenze sul mondo, compie un'azione, il mondo cambia e molte delle credenze che essa aveva devono essere riviste o aggiornate. Come? Non è possibile che noi percepiamo e notiamo tutti i cambiamenti (in primo luogo, perché molti dei cambiamenti che sappiamo che avvengono, non avvengono nel nostro campo percettivo), e quindi non possiamo affidarci interamente agli input percettivi per rivedere le nostre credenze. Dobbiamo avere modi interni per aggiornare le nostre credenze, che colmeranno le lacune e manterranno il nostro modello interno (la totalità delle nostre credenze) in corrispondenza più o meno fedele con il mondo. (Dennett 1978: 92)

In sostanza, il problema si presenta quando un agente con capacità cognitive deve agire efficacemente nel mondo e riuscire a comprendere quali siano le caratteristiche più importanti da tenere in considerazione affinché l'azione che si ha intenzione di svolgere possa essere portata a compimento. Per comprendere meglio l'oggetto del problema, si pensi, ad esempio, di chiedere al nostro partner di prepararci un panino al formaggio. Per far sì che riesca a soddisfare la nostra fame, il nostro partner dovrà sapere cosa sia un panino, cosa sia il formaggio, se servono altri ingredienti oltre al pane e al formaggio, che magari sarebbe meglio grigliarlo, che per grigliarlo servono determinati utensili e dove si trovano gli ingredienti e gli strumenti in cucina. Queste sono tutte informazioni riguardo il contesto. In seguito, il nostro partner deve essere a conoscenza di una serie di nozioni implicite: come tagliare il pane affinché possa contenere gli ingredienti del panino, e saper dunque maneggiare un coltello, deve sapere che il formaggio si dà in molte forme (cremoso, solido, semi-stagionato), che può essere spalmato, tagliato o sciolto (o una combinazione di queste azioni), maneggiare quindi il formaggio in modo appropriato, e assemblare il panino in modo che resti insieme mentre lo mangio. Poi c'è bisogno che il nostro partner sappia riconoscere gli ingredienti anche una volta che il loro stato è cambiato: il pane passa da essere intero a essere diviso in due, il formaggio si modifica diventando a fettine, e è importante saperli maneggiare anche dopo che hanno cambiato stato (cambiamento che magari implica un diverso trattamento). Infine, c'è bisogno che il nostro partner sappia anche improvvisare: se io ho chiesto espressamente un panino al formaggio, ma il mio partner scopre che ne siamo sprovvisti perché sono giorni che non facciamo la spesa, cosa può fare per accontentarmi? Magari può orientarsi su un altro ingrediente, come il prosciutto, sperando che un panino con il prosciutto mi andrà comunque bene (perché è pur sempre un panino), oppure potrà

scegliere di uscire di casa e andare a fare la spesa se sa che non sono molto incline ai cambiamenti di piano sui miei pasti.

Di fatto, anche se un compito del genere può sembrare semplicissimo, questo implica una serie molto ampia (quasi infinita) di specificazioni rispetto al modo in cui portiamo avanti le nostre azioni e ai diversi tentativi che facciamo per affrontare le contingenze che ci troviamo davanti quando le svolgiamo. Ovviamente, nel campo dell'intelligenza artificiale questo problema risulta essere di vitale importanza, soprattutto nei domini in cui molte nozioni e concetti vengono dati per scontati, cosa che ogni essere umano fa per la maggior parte del tempo. Abbiamo già visto applicazioni di questo problema per SHRDLU e SHAKEY: come fare a rendere capace SHRDLU di riconoscere i blocchi una volta che il loro stato cambia (ad esempio, una piramide da essere sopra a un cubo passa a essere nella scatola)? Come consentire a SHAKEY di trovare opzioni alternative intelligenti e in tempo reale quando la pianificazione di partenza risulta inattuabile?

Il fatto che gli sviluppatori di questi programmi e robot siano incappati nel *frame problem* è uno dei motivi principali per cui la loro impresa è stata considerata anche un'attestazione delle inefficienze del metodo GOFAI. Soprattutto in robotica, il problema si mostra in tutta la sua complessità e insormontabilità: dal momento in cui si tolgono dai micromondi in cui tutte le caratteristiche e le possibili azioni erano specificate, i robot devono avere a che fare con numerose altre variabili, legate alla materialità e alla contingenza. Anche nel caso di SHAKEY, che si muoveva in un mondo giocattolo costruito a posta per lui, il problema della materialità e dell'imprevedibilità del mondo faceva esplodere la possibilità di riuscire ad agire all'interno di quel mondo avendo a disposizione solamente un insieme di simboli preimpostati che lo descrivevano e una serie di algoritmi per manipolarli. Nello scenario che si prospettava in quel momento, in cui si voleva che i robot fossero capaci di interagire nel mondo 'reale' (Moravec 1980), il problema creava difficoltà non banali e di non facile risoluzione<sup>69</sup>.

Alcuni tentativi per ovviare al problema sono stati provati all'interno dello stesso dominio in cui il problema era stato posto. In linea con la tradizione della GOFAI, alcuni esponenti dell'AI classica hanno costruito una serie di strumenti teorici utili alla formalizzazione della conoscenza implicita e contestuale. Senza dubbio, il tentativo più celebre è quello portato avanti da Minsky, che elabora la cosiddetta teoria dei *frame* proprio per trovare una soluzione al *frame problem*. Nell'idea di Minsky,

*A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame.*

---

<sup>69</sup> La fantasiosa e divertente storia narrata da Dennett (1984: 129) riguardo ai robot R1 e successivi rende bene l'idea del problema dei *frame* in robotica. Per elaborazioni del *frame problem* in robotica nell'ambito della filosofia della mente e delle scienze cognitive computazionaliste: Pylyshyn 1987; Ford e Pylyshyn 1996.

Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed. (Minsky 1974: 111-112)

Usando questa nozione, l'obiettivo di Minsky era quello di mappare tutta la conoscenza riguardo una determinata situazione per consentire al sistema artificiale, computer o robot, di riuscire a trovare il giusto *frame* che selezioni il quadro interpretativo entro cui inscrivere la situazione in cui si trova. In sostanza, l'approccio avanzato da Minsky suggeriva che se si fosse stati capaci di mappare e formalizzare l'adeguato numero di dati nel giusto tipo di conoscenza e di gerarchizzarlo in termini di importanza relativa a una situazione, i sistemi artificiali sarebbero stati capaci di avere a che fare con il "senso comune" allo stesso modo degli umani.

I tentativi di implementazioni della teoria dei *frame* elaborata da Minsky sono stati numerosi (Boden 2006: 748), ma il presupposto che guidava lo sviluppo di questa idea era strettamente legato a quello che Dreyfus aveva indentificato come l'assunto epistemologico: si doveva presupporre che il mondo fosse scomponibile atomicamente e descrivibile in modo formale in tutte le sue parti per poter arrivare a pensare di poter risolvere il *frame problem* in questo modo. Dreyfus, d'altra parte, era molto più scettico. In sostanza, infatti, come sottolineato anche da Haugeland, il *frame problem* si caratterizza come un problema riguardante il fatto che le inferenze che vengono fatte nel contesto delle azioni che svolgiamo hanno natura olistica: «The so-called frame problem is how to 'notice' salient [inferences] without having to eliminate all of the other possibilities that might conceivably have occurred had the facts somehow been different» (Haugeland 1985: 204). Si può immaginare che, a partire dalla posizione fenomenologica heideggeriana, la critica a questo approccio sarebbe stata immediata. Dreyfus, infatti, pone l'accento sul fatto che il problema si distingue in termini di rilevanza:

It seemed to me, however, that the deep problem wasn't storing millions of facts; it was knowing which facts were relevant in any given situation. [...] But a system of frames isn't *in* a situation, so in order to select the possibly relevant facts in the current situation one would need frames for recognizing situations like birthday parties, and for telling them from other situations such as ordering in a restaurant. But how, I wondered, could the computer select from the supposed millions of frames in its memory the relevant frame for selecting the birthday party frame as the relevant frame, so as to see the current relevance of, say, and exchange of gifts rather than money? (Dreyfus 2008: 332-333)

Il necessario conseguente di questa impostazione, continua Dreyfus, è che ci sia un regresso a infinito della selezione dei *frame*. Quando si può capire che ci si deve fermare? Se la teoria dei frame costituisce un tentativo di formalizzare la capacità di buon senso nel discernimento della giusta postura interpretativa di fronte a una situazione data, il fallimento è dato dal fatto che viene postulata la stessa facoltà che si cerca di spiegare o formalizzare per far sì che il processo abbia luogo. Appoggiandosi alle riflessioni wittgensteiniane in un altro saggio scritto con suo fratello Stuart, Dreyfus mostra, in sostanza, come il tentativo di formalizzazione attuato con la teoria dei frame si

configuri come la prova che ci sono parti dell'esperienza umana che non possono essere formalizzate tramite questo approccio.

When we try to find the ultimate context-free, purpose-free elements, as we must if we are going to find the primitive symbols to feed a computer, we are in effect trying to free aspects of our experience of just that pragmatic organization which makes it possible to use them intelligently in coping with everyday problems. (Dreyfus e Dreyfus 1988: 26)

Ciò dimostra, secondo Dreyfus, che «the frame problem wasn't just a problem but was a sign that something was seriously wrong with the whole approach» (Ivi). Teorie come quella appena esposta o il confinamento dei programmi in micromondi costruiti appositamente potevano costituire tentativi di temporanea soddisfazione dei principi, ma era la teoria stessa che si dimostrava essere inappropriata per raggiungere gli scopi che si era prefissata. Il modo attraverso cui noi esseri umani arriviamo a saper esercitare certe scelte interpretative anche in modo implicito passa attraverso un'esperienza di apprendimento strettamente dipendente dal modo in cui siamo al mondo e da cosa viene indicato come rilevante nel contesto biologico, sociale, culturale che ci troviamo ad occupare (Dreyfus 1972, 1991, 1992, 2008).

L'interpretazione di Dreyfus conduce non solo a ridimensionare il ruolo esplicativo delle tecniche sviluppate nel dominio di riconoscimento di oggetti o di comprensione del significato delle azioni di sistemi artificiali come i robot, ma comporta una riflessione generale sull'impostazione teorica tutta da cui l'impresa dell'intelligenza artificiale ha preso avvio.

But if robots for processing nonformal information must be [...] entirely different from present digital computers, what can now be done? (Dreyfus 1972: 216)

Cosa fare, dunque? Dal momento che il *frame problem* sembra insormontabile se posto in questi termini, una delle soluzioni suggerite è quella di provare a cambiare la teoria (o la filosofia) che supporta l'impresa della costruzione di artefatti intelligenti. A partire dalle considerazioni fatte da Dreyfus, Weizenbaum e Winograd e Flores, questo è ciò che cercherà di fare la corrente alternativa alla GOFAI nelle scienze dell'artificiale, cercando di sostituire i principi rappresentazionisti, simbolici e computazionalisti con soluzioni provenienti dalle teorie viste in questa sezione.

### 2.3.2 *Nouvelle AI*

Negli anni '80, le critiche formulate da Weizenbaum, Winograd e Dreyfus erano già in circolazione da quasi vent'anni, ma risultati significativi nel campo della robotica non erano ancora stati registrati. Infatti, la maggior parte delle ricerche nel campo della robotica si basava ancora sui processi di pianificazione simbolica e di costruzione di micromondi (Brady 1985; Minsky 1985), ma si possono contare diverse testimonianze riguardanti lo sviluppo di un approccio più *situato*, che si rifaceva agli esempi di robotica costruiti da Grey Walter decenni prima (Moravec 1999: 15-49; cfr. §1.2.3). In

particolare, contributi teorici come quelli di Valentino Braitenberg (1984) avevano contribuito enormemente a rifocalizzare l'attenzione dei costruttori di robot su meccanismi di replicazione più basilari, molto più vicini a quelli sviluppati durante la prima cibernetica che alle ultime attestazioni provenienti dall'intelligenza artificiale (Datteri 2013): le tartarughe di Grey Walter sembravano comportarsi in modo più intelligente nell'ambiente rispetto a SHAKEY, anche se la tecnologia che permetteva loro di muoversi era di gran lunga meno avanzata di quella implementata nei robot successivi. In sostanza, i tentativi di rielaborazione dell'approccio robotico e i discorsi provenienti dalla controcultura di cui abbiamo parlato indirizzarono la ricerca non più solamente verso modalità di implementazione di ragionamenti astratti, cercando quindi di replicare l'intelligenza dall'interno, ma condussero a un rinnovato interesse riguardo il comportamento esibito degli agenti intelligenti che, visti dal punto di vista dell'osservatore, potevano essere in grado di comportarsi in modi anche imprevedibili.

È intorno a questa nuova corrente di pensiero che si raccolgono le ricerche della nuova robotica, che nasce negli anni '80 prendendo il nome di *Nouvelle AI* (Brooks 1986, 1989, 1991a, 1991b, 1991c). È possibile raccogliere in un unico quadro descrittivo i punti cardine di questi nuovi approcci alla progettazione della robotica seguendo quattro principi fondamentali:

1. Per costruire robot 'intelligenti', si deve procedere a ingegnerizzarli, non a programmarli: invece che cercare di costruire l'apparato astratto di regole e simboli, adottando quindi un approccio *top-down*, si deve dotare il robot della capacità di sviluppare autonomamente il proprio grado di sensibilità dato dalla sua struttura "corporea" e fare in modo che sfrutti le proprie capacità sensorimotorie per esplorare il proprio ambiente (approccio *bottom-up*). Solo così sarà possibile arrivare a risultati più performanti, affidabili ed ecologicamente validi.
2. Non è scontato che si possa arrivare a comprendere e a modellare l'intelligenza umana, ma, se questo è possibile, sarà solamente l'ultimo passo in un processo in cui si saranno modellizzate e comprese forme più 'semplici' di intelligenza animale, come quella degli insetti.
3. Lo sforzo dell'intelligenza artificiale deve essere diretto alla modellazione di sistemi sensorimotori completi: invece che scomporre i compiti in diverse sub-aree dell'AI, il comportamento degli agenti intelligenti deve essere compreso come il risultato dell'insieme olistico della struttura corporea immersa in un ambiente, ovvero ciò che è stato definito "The whole iguana" da Dennett (1978).
4. La biologia deve essere la disciplina con cui portare avanti il confronto, sostituendo quindi la logica come paradigma direttore dell'indagine.

Solitamente, la corrente a cui si fa maggiormente riferimento è quella fondata dal già citato Rodney Brooks (1986, 2002; §1.4.2.1), sicuramente quella che ha goduto di una fama più ampia, ma non è l'unica. Infatti, parallelamente alle ricerche di Brooks e colleghi nel campo nella neonata *behavior-based robotics*, altri due filoni di ricerca contribuiscono a stabilire un nuovo paradigma per la robotica, che dopo questa vera e propria rivoluzione non poteva continuare a basarsi esclusivamente sul concetto di pianificazione.

Da una parte, l'approccio denominato *situated automata theory* nasce già all'inizio degli anni '80 nel campo della linguistica computazionale e si sviluppa in robotica grazie al lavoro, tra gli altri, di Leslie Kaelbling (1991). Il focus di ricerca era il legame tra l'agente e l'ambiente in cui si muoveva: invece che progettare il bagaglio di conoscenze simboliche e le regole di manipolazione come era consueto nell'approccio *top-down*, lo scopo era quello di garantire al robot di sapere quali fossero le circostanze in cui avrebbe dovuto svolgere il task, costituendo dunque un approccio *bottom-up* (Rosenschein e Kaelbling 1995). Il traguardo più importante raggiunto da questo approccio è la progettazione del robot Flakey che, a discapito dell'assonanza tra i nomi, era progettato in modo completamente diverso da SHAKEY. Le capacità di Flakey, infatti, non erano garantite dall'implementazione di un programma specifico, ma principalmente da circuiti hardware attraverso cui riusciva a captare informazioni riguardo l'ambiente in cui si trovava e a svolgere azioni solo nel momento in cui le circostanze lo permettessero. Nonostante il sostanziale miglioramento delle performance in relazione a quelle dei suoi predecessori, Flakey non era costruito abbandonando completamente l'account simbolico e rappresentazionalista classico (Ibid.: 150).

Alternativamente, l'Intelligenza Artificiale Interazionista (*Interactionist AI*) si concentrava principalmente sull'obiettivo di eliminare totalmente l'uso delle rappresentazioni interne nella costruzione della cognizione robotica, favorendo invece un approccio heideggeriano all'intelligenza artificiale (anche detto *Heideggerian AI*, Kiverstein 2012). Un noto esempio di questa linea di ricerca è la proposta di Philip Agre, il quale ribaltò totalmente lo sguardo nei confronti del problema della modellazione della cognizione e del comportamento:

Faced within an empirical phenomenon to explain, our first explanatory recourse should be to dynamics, not to machinery. Faced with a technical problem to solve, our engineering should begin with dynamics, not with machinery. (Agre 1988: 27)

Come sottolineato da Boden (2006: 1033), il concetto di 'dinamica' risulta centrale nell'approccio di Agre, fortemente influenzato dalle teorie di Dreyfus, dal momento che comporta uno spostamento di attenzione dalla modellazione interna dell'architettura dell'agente al comportamento esibito dallo stesso in interazione con il suo ambiente. Uno degli esempi più interessanti sviluppati da Agre (e Chapman 1987) è il programma *Pengi*, un videogioco con protagonista un pinguino che non si basava su alcun modello interno per interagire nel suo ambiente e con gli oggetti che ne facevano parte, ma

utilizzava rappresentazioni temporanee (deittiche) legate all'attività che stava svolgendo e ancorate agli oggetti che aveva 'a portata di mano'.

Tuttavia, come si può notare dagli esempi, nessuna delle tecnologie sviluppate da queste correnti poteva essere considerata rivoluzionaria rispetto allo stato dell'arte della robotica a loro contemporanea. L'estrema originalità delle tesi avanzate, rispetto alla centralità dell'interazione tra agente *embodied* e mondo o all'eliminazione delle rappresentazioni interne, rendeva necessaria una forte innovazione non solo dal punto di vista teorico, che comprendeva un apparato capace di intrecciare percezione e azione, cognizione e mondo, ma anche materialmente legato alla tecnologia di supporto. Ciò fu possibile solamente con il nuovo approccio alla robotica elaborato da Rodney Brooks al MIT a partire dalla fine degli anni '80, che lui stesso chiamerà *behavior-based robotics*.

### 2.3.2.1 Behavior-based robotics

Rodney Brooks, ricercatore in robotica al MIT in quegli anni, arrivò a sviluppare una nuova modalità di costruire robot che sembrava fornire una soluzione a coloro che cercavano un'alternativa agli approcci tradizionali dell'intelligenza artificiale. Infatti, nella elaborazione per una nuova robotica, Brooks convoglia tutte le rivoluzioni teoriche che avevano maturato nel corso del periodo precedente: il pensiero alternativo si manifesta nei numerosi richiami a Dreyfus (Brooks 1986, 1991a, 2002), a Winograd e Flores (Brooks 1989, 1991a, 1991b) e a Weizenbaum (Brooks 2002), e ai contributi recenti dei nuovi approcci interazionisti e situati di Agre e Kaelbling (Brooks 1991a, 1991b, 1991c), disseminati nelle opere di Brooks, considerate il manifesto della Nouvelle AI e prese a canone di una modalità alternativa di incorporare la cognizione in meccanismi artificiali.

L'approccio di Brooks della *behavior-based robotics* è inquadrabile attraverso una serie di principi epistemologici in cui viene chiaramente rappresentato il distacco dall'approccio tradizionale e si intravedono le radici dei presupposti teorici che instraderanno la costruzione dei nuovi robot.

In particolare, i principi epistemologici che si delineano riguardano tre ambiti:

- *Ambito evolutivo: il comportamento intelligente deve essere modellato a partire dalle sue componenti basilari* (Brooks 1991a). In questo principio, Brooks raccoglie una lezione già espressa da Moravec (1988) circa il fatto di rivolgere lo sguardo non al modo in cui l'intelligenza si manifesta oggi, ma al modo in cui l'evoluzione ha permesso l'intelligenza nel suo sviluppo. La chiave, sottolinea Brooks, è capire come capacità di altissimo livello possedute ora dagli esseri umani, come il linguaggio, il problem-solving e la pianificazione – tutte caratteristiche oggetto di interesse dell'approccio tradizionale all'intelligenza artificiale – si siano evolute in tempi relativamente recenti in termini evoluzionistici e si siano appoggiate per emergere a capacità molto più radicate nell'organismo, come la capacità di movimento e



di percezione dell'ambiente in cui siamo inseriti. È per questo che, seguendo il paradosso di Moravec, è molto più complesso riprodurre capacità che al nostro occhio sembrano più semplici: le capacità che legano percezione e azione, che hanno richiesto all'evoluzione più tempo per essere sviluppate, sono alla base stessa di tutta l'intelligenza umana. Allora, è su queste capacità, quelle legate alla corporeità dell'organismo e al suo strettissimo rapporto con l'ambiente, che deve concentrarsi l'attenzione e lo sforzo di imitazione.

- *Ambito biologico-ingegneristico: la costruzione delle creature artificiali deve riguardare necessariamente l'agente nel suo complesso, e non solamente alcune sue parti* (Brooks 1989). A partire da un commento formulato da Dennett (1978), Brooks prende spunto per costruire il suo approccio dell'*iguana intero*. Nel suo articolo, Dennett aveva ipotizzato che costruire un sistema artificiale che riproducesse le capacità di un essere umano nella sua interezza sarebbe stato molto complicato e non attuabile come proposito a breve termine. La scommessa che lancia è quindi quella di cercare di imitare organismi più semplici, come le stelle marine o le tartarughe, magari un'iguana. Brooks raccoglie la sfida e articola il suo saggio (Brooks 1989) proprio sull'intenzione di voler simulare nel complesso il comportamento di un organismo più semplice di quello umano, cioè quello degli insetti: l'accento è posto sul modo in cui l'organismo dimostra la sua intelligenza tramite il suo comportamento nell'ambiente all'interno del quale è inserito, tralasciando la riproduzione fedele dei meccanismi biologici che lo generano. Riprendendo una lezione elaborata decenni prima da Herbert Simon (1969, Boden 2006: 429-430), organismi come le formiche dimostrano un comportamento molto complesso se osservate camminare sulla sabbia, ma i meccanismi che permettono quel comportamento non lo sono: non ci sono capacità di pianificazione, dal momento che la formica non ha un cervello; non c'è bisogno che la formica conosca tutto il suo mondo, l'importante è che riesca a individuare gli ostacoli quando le si parano davanti, cambi direzione e non perda la capacità di muoversi nel suo ambiente; anche se il suo andamento sembra traballante e tentativo, dimostra una notevole capacità adattiva nel suo ambiente.
- *Ambito delle scienze cognitive: l'ipotesi del sistema fisico di simboli deve essere sostituita con l'ipotesi dell'ancoramento fisico (Physical Grounding Hypothesis)* (Brooks 1990). A partire da una critica serrata all'ipotesi di Newell e Simon, l'obiettivo polemico si allarga fino ad includere tutte le scienze cognitive computazionaliste simboliche e le loro realizzazioni pratiche nel settore dell'intelligenza artificiale. La soluzione è allora quella di sostituire la nozione di sistema simbolico con quella di ancoramento fisico, in cui la creatura artificiale è saldamente situata nel suo mondo, non in una riproduzione interna di esso:

Nouvelle AI is based on the physical grounding hypothesis. This hypothesis states that to build a system that is intelligent it is necessary to have its representations grounded in the physical world. [...] Accepting the physical grounding hypothesis as a basis for research entails building systems in a bottom up manner. (Brooks 1990: 5)

È allora dalle capacità di interazione con l'ambiente e dalle modalità in cui l'agente mette in pratica il suo accoppiamento con il mondo che è possibile sviluppare forme di intelligenza di alto livello. Il processo contrario, quello tipico della GOFAI in cui le modalità di interazione materiale con l'ambiente sono modellizzate tramite proposizioni nella logica formale che agiscono su un mondo astratto interno, non porterà mai alla costruzione di un agente intelligente al livello dell'intelligenza umana. Per Brooks, molte delle ricerche svolte negli ultimi 25 anni si sono rivelate irrilevanti per la costruzione di esseri artificiali nel mondo reale (Brooks 1989: 1).

L'adozione di questo piano di intenti si realizza attraverso l'elaborazione di una serie di principi, in cui l'apparato epistemologico viene declinato operativamente e si delinea l'impianto teorico che funge da linea direttrice per la robotica basata sul comportamento. In primo luogo, ripartendo dalle critiche ai robot della GOFAI che abbiamo visto in §2.3.1, Brooks si schiera a favore di un approccio che non si basa sulla modellazione interna del mondo, ma che prende il mondo come il suo miglior modello (Brooks 1991a: 15): «the world is its own best model» diventa lo slogan attraverso cui identificare l'approccio anti-rappresentazionalista della robotica situata. La dichiarazione della maggiore appropriatezza del mondo reale nella funzione di modello per le creature artificiali si manifesta anche nella risposta che Brooks (1989) dà a Dennett (1978) nell'approccio dell'*"iguana intero"*: nella sua scommessa per la modellazione di un agente completo, Dennett non solo ipotizza che ci sia bisogno di astrarre la struttura del comportamento dell'agente, ma anche di costruire il modello del mondo all'interno del quale l'agente deve agire. Per questo, Dennett scommette sulla costruzione di un'iguana marziano, in cui l'ambiente sia costruito appositamente per facilitare l'interazione. Brooks sottolinea esattamente questo punto:

[Dennett's] mistake was to suggest simulating the environment. It requires a great deal of work to simulate a rich enough environment to make the problem for the creature realistic; more work, I claim, than in simulating the creature itself. (Brooks 1989: 2)

A differenza dei mondi virtuali alla base dei ragionamenti e delle pianificazioni di SHAKEY e SHRDLU, il mondo reale e materiale è sempre dettagliato, non ha bisogno di aggiornamenti, e soprattutto non deve essere costruito. È assurdo, allora, non prendere in considerazione il mondo reale così com'è, lo stesso in cui abitiamo noi, per costruire organismi che dimostrino un grado alto di intelligenza nell'interazione. La *situatedness* è quindi il primo principio su cui basare la costruzione delle creature artificiali.

Questo principio comporta anche l'assunzione di un secondo assunto, quello dell'*embodiment* dell'organismo.

There are two reasons that embodiment of intelligent systems is critical. First, only an embodied intelligent agent is fully validated as one that can deal with the real world. Second, only through a physical grounding can any symbolic or other system find a place to bottom out, and give 'meaning' to the processing going on within the system. (Brooks 1991a: 15)

Il robot deve essere capace di *sentire* con il suo apparato sensibile le caratteristiche del mondo in cui deve agire, così da riuscire a reagire efficacemente in base alle potenzialità che il suo stesso apparato corporeo – il suo hardware – gli consente. Solo così, secondo Brooks, è possibile uscire dalla spirale del regresso ad infinito aperta dal *frame problem*, dal momento che senza partecipazione e percezione incarnate del mondo non c'è significato per l'agente: «the world grounds regress» (Ibid.: 16).

In questo quadro, allora, il concetto di intelligenza si modifica e passa dall'essere esclusivamente la capacità di ragionamento astratto e complesso, a manifestarsi prima di tutto nel rapporto che l'agente intrattiene con il mondo:

It is hard to draw the line at what is intelligence, and what is environmental interaction. In a sense it does not really matter which is which, as all intelligent systems must be situated in some world or other if they are to be useful entities. (Brooks 1991a: 16)

Si potrebbe mai dire che un elefante non sia intelligente perché non è capace di giocare a scacchi (Brooks 1990)? Nell'articolazione di questo paradosso, Brooks fa notare che il giudizio riguardo all'intelligenza di un organismo è strettamente dipendente dal modo in cui viene contestualizzata l'azione svolta dall'organismo stesso, messa in relazione alle capacità sviluppate all'organismo e al modo in cui si dimostra un grado di efficacia nel mondo. Ciò che è cruciale è dunque rendere pertinente il ruolo dell'intelligenza nel comportamento della creatura e individuarne le caratteristiche basilari affinché sia possibile imitarne i risultati. Come riprodurre, quindi, comportamenti intelligenti?

Il modo attraverso cui l'intelligenza si manifesta è dato nell'agente dalle sue capacità di auto-organizzazione interna, la cui complessità è data dai processi evolutivi che hanno garantito l'*emergenza* della qualità a partire dall'interazione delle componenti più basilari dell'organismo. Allo stesso modo, dunque, il robot *behavior-based* deve cercare di replicare le manifestazioni di intelligenza degli organismi semplici come gli insetti, mettendo in interazione componenti più basilari, da cui emergerà poi il comportamento generale. Facendo riferimento agli studi a lui contemporanei sull'auto-organizzazione in campo biologico (Nicolis e Prigogine 1985) e artificiale (Steels 1990), Brooks adotta un approccio secondo cui non sempre un comportamento intelligente è riconducibile in un rapporto diretto all'interazione di determinate componenti organizzazionali, ma è

spesso il risultato di una interazione indiretta di più componenti primitive sia tra di loro, sia con l'ambiente. In sostanza, ciò che viene identificato come intelligente non è altro che il risultato in superficie di una serie di interazioni che avvengono su più livelli e su più fronti:

Intelligence can only be determined by the total behavior of the system and how that behavior appears in relation to the environment. (Brooks 1991a: 16)

L'idea sostanziale, che vediamo essere ripresa dalla tradizione della seconda cibernetica risalente a von Foerster e ai suoi allievi Maturana e Varela, è che l'intelligenza, come altre proprietà emergenti dell'organismo, è *nell'occhio dell'osservatore*: ciò non significa negare l'importanza dell'organizzazione biologica o funzionale delle componenti primitive, ma comporta la presa in considerazione dello sguardo osservante sul processo di individuazione delle capacità stesse dimostrate. La chiave è costruire una creatura che dimostri agli occhi degli osservatori capacità adattive rispetto al suo ambiente.

Ora, la domanda che sorge spontanea è: è possibile implementare tutti i principi alla base della nuova epistemologia della nouvelle AI in una struttura artificiale che li rispetti? Forse, il successo che ha investito questo paradigma in robotica è derivato proprio dal fatto che le creature costruite da Brooks sembravano incarnare perfettamente i principi enucleati nei suoi scritti teorici. La chiave per realizzare questi processi consisteva nel nuovo tipo di architettura inventata da Brooks, ovvero la *subsumption architecture*.

Le caratteristiche principali di questo nuovo tipo di architetture, definite anche architetture in parallelo, consistevano nella composizione del comportamento e nella disposizione del sistema di controllo attraverso degli strati, o *layer* (Brooks 1989: 433), per cui: i) il problema veniva decomposto in base a comportamenti orientati al soddisfacimento di un solo task, che potevano lavorare in parallelo; ii) ogni comportamento è costruito a partire da semplici macchine a stati finiti<sup>70</sup> che possono comunicare tra di loro e sono coordinate in modo asincronico, ma per cui ad ogni input corrisponde un output; iii) viene modellizzato un comportamento alla volta, sviluppato per essere utile al raggiungimento di uno scopo, senza mai essere modificato o reso più complesso: per aggiungere capacità si devono aggiungere ulteriori *layer* di controllo; iv) tutta la capacità computazionale viene spostata a bordo del robot, senza quindi delegarne alcuna parte a un computer al di fuori di questo<sup>71</sup>.

---

<sup>70</sup> Per la definizione di macchina a stati finite, riportiamo un esempio che può renderne chiara la struttura: «The concept of state is very familiar. Consider a toaster: initially, the toaster is in the off state; when you push the lever down, a transition is made to the on state and the heating elements are turned on; finally, when the timer expires, a transition is made back to the off state and the heating elements are turned off.

A finite state machine (FSM)<sup>1</sup> consists of a set of states  $s_i$  and a set of transitions between pairs of states  $s_i, s_j$ . A transition is labeled condition/action: a condition that causes the transition to be taken and an action that is performed when the transition is taken». (Ben-Ari e Mondada 2018: 55)

<sup>71</sup> Senza dubbio, una buona parte di questo risultato è dovuta all'invenzione dei microprocessori, che miniaturizzavano la capacità dei sistemi fino ad allora usati in robotica, senza i quali sarebbe stato impossibile implementare la computazione a bordo (Bernstein 1978/1982; Brooks 2002).

La prime conseguenze dirette di questo approccio sono che, da una parte, il robot può essere testato direttamente nell'ambiente, senza costruire alcun mondo semplificato per lo scopo di farlo performare; dall'altra, non è individuabile alcun centro di controllo in cui l'informazione viene convogliata e gestita, nessuna struttura centralizzata dei dati, e nessun piano globale per il comportamento da dimostrare. In poche parole, «there is no way to tell the robot to do something» (Ibid.: 435). In sostanza, la modellazione del comportamento avveniva costruendo strati di controllo semi-indipendenti ognuno dei quali corrisponde a un livello di competenza da dimostrare per raggiungere il task: aumentando il numero di strati, e quindi di competenze, il comportamento poteva essere complessificato.

Il nucleo dell'architettura a sussunzione consiste nel permettere ai vari strati di performare indipendentemente uno dall'altro, facendo però tutti riferimento ai dati raccolti dai sensori e riportando le proprie elaborazioni agli stessi attuatori: la struttura di base è sempre costituita da un livello 0, alla base del comportamento, che garantisce lo svolgimento del task di livello più basso, come evitare gli ostacoli che la creatura trova sul proprio cammino; di seguito, ogni livello viene aggiunto in modo tale che sia possibile sfruttare lo stesso flusso di dati in entrata, ad esempio vagando in un ambiente, ma senza che il livello inferiore sia a conoscenza del livello superiore, quando invece il livello superiore può all'occorrenza inibire qualche competenza di livello inferiore.

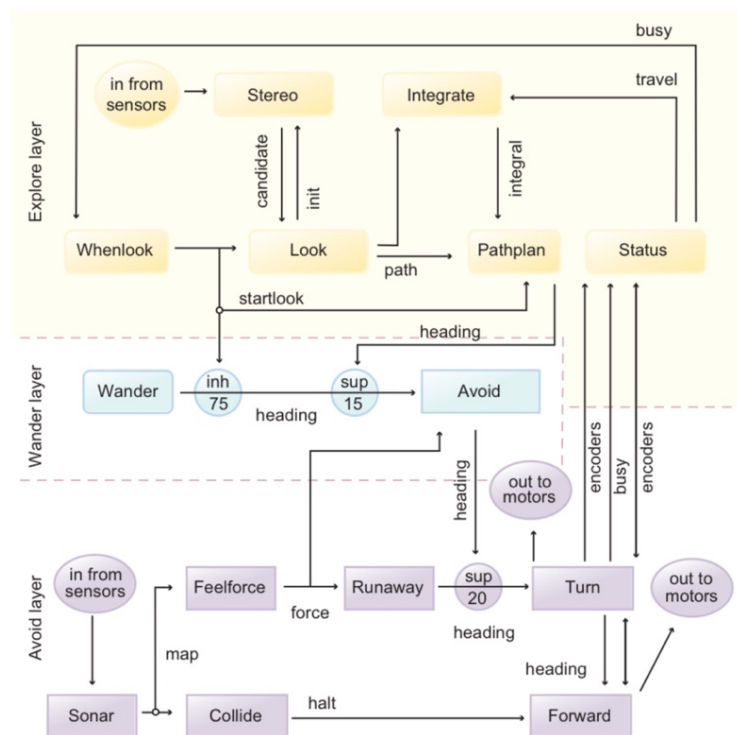


Fig. 1.3 Una rappresentazione della prima *subsumption architecture*, implementata nel robot Allen (gli strati comunicano attraverso meccanismi di inibizione [inh] e soppressione [sup]). (Bermudez 2020: 426, adattato da Brooks 1991c)

Il complesso dei vari livelli può riuscire a sviluppare comportamenti più complessi rispetto a quelli realizzati tramite controllo centralizzato e rappresentazione simbolica: di fatti, il robot di Brooks vagavano per la stanza evitando ostacoli, facendo affidamento principalmente ai dati provenienti dall'ambiente e appoggiandosi alla capacità di evitare gli ostacoli modellizzata nello strato di livello zero.

Come notato in più sedi da Brooks stesso, questo rendeva il comportamento del robot molto più robusto, evitando quindi la fragilità delle architetture tradizionali che dovevano essere aggiornate ogni qualvolta si verificasse un impedimento dall'esterno. I robot costruiti da Brooks e dal suo gruppo, come Allen, Herbert, Tom e Jerry, Genghis, Toto o Seymour (Brooks 1990) erano tutti basati su questi principi e riuscivano a dimostrare capacità di interazione con l'ambiente mai raggiunte fino a quel momento: riuscivano a restare in attività per ore, recuperando quando si trovavano di fronte ad alcune difficoltà, eseguendo le azioni in tempo reale senza l'intervento umano, cercando di trovare il percorso libero da ostacoli invece di averli tutti presenti – come facevano Genghis o Allen – o raccogliendo lattine e buttarle nella spazzatura – come era capace di fare Herbert (Brooks 1989: 451).

Ora, riesce il modello elaborato da Brooks a soddisfare i principi epistemologici e teorici sui cui dichiara basarsi la sua impresa ingegneristica? Per rispondere a questa domanda, bisogna tenere conto che i principi enunciati da Brooks nei suoi articoli più teorici vanno in ogni caso visti come un tentativo di modellazione, da lui stesso dichiarato, che si affida a principi epistemologici ben precisi, e devono essere messi in correlazione alle rispettive capacità di *operativizzazione* rese possibili dalla tecnologia. Rimandando dunque le considerazioni generali rispetto all'efficacia delle creature di Brooks ai paragrafi che seguiranno, se assumiamo il punto di vista dell'osservatore e ci poniamo di fronte ai robot che interagiscono in tempo reale e in modo efficace con il proprio ambiente per un prolungato lasso di tempo, è chiaro che l'impressione suscitata si sia rivelata di gran lunga migliore rispetto all'osservazione delle performance dei robot precedenti. La reazione è attestata dalla maggior parte degli esperti del tempo nel campo della robotica e dell'intelligenza artificiale<sup>72</sup> (Boden 2006: 1300-1302). A partire da questi riconoscimenti, Brooks abbandona l'operazione di riproduzione degli insetti e dei comportamenti di livello più basso, considerati dimostrati i risultati desiderati e apertasi una nuova linea di ricerca sulla base delle sue realizzazioni precedenti. Insieme ai suoi collaboratori, allora, Brooks si lancia nell'impresa, forse troppo ottimistica fin dal principio, di replicare forme molto più complesse di comportamento nel progetto del robot *Cog*.

---

<sup>72</sup> D'altra parte, come succede non di rado ai progetti che più si distaccano da un'impostazione tradizionale per percorrere una via non ordinaria, anche i progetti di Brooks (e Brooks in persona) sono stati oggetto di aspre critiche (Hsu 2023: 96).

### 2.3.2.2 Cognitive Social Robotics

A seguito delle importanti innovazioni introdotte nel settore della robotica dall'approccio della Nouvelle AI, Brooks fonda il progetto *Cog* al MIT nel 1993 con un duplice obiettivo: i) nel campo della robotica, costruire un prototipo di robot autonomo abile e flessibile, impiegabile per diversi scopi; ii) nel campo delle scienze cognitive, comprendere la struttura della cognizione umana (Brooks e Stein 1994; Brooks et al. 1998). Come è facile intuire, gli obiettivi che guidavano l'indagine erano di gran lunga più complessi di quelli che avevano dato inizio alla svolta della *behavior-based robotics* iniziata circa quindici anni prima, ma l'accelerazione era data da un numero di fattori anche di natura molto differente tra di loro. Come sottolineato da Brooks in una recente intervista:

[...] people were very surprised when I went from insects to humanoids. And I have got two stories about it. One, it had taken so many years to do the insects and I looked at how much time I had left in life if I was going to do little lizards and then mammals, and then monkeys. You know, I did not have enough time to do all those cycles, so maybe I should go straight for it. That was one aspect. The other aspect was on January 12th, 1992, I had my grad students over and we watched *2001, A Space Odyssey*. That day was the birth of HAL. And we are like, "Where the hell is HAL? HAL is not here. What, what?" So, that was actually what prompted me to think, "Okay, let us try something. Let us do a big jump and let us try and look at human level intelligence. What can we do there using the techniques we have developed over the last ten years, what can we do?" And that is where the humanoid robots came from, and Cog was the first one of them. (Hsu 2023: 99)

La spinta verso la costruzione di robot umanoidi, che quindi cercassero di implementare non più architetture che facessero emergere competenze semplici come quelli di un insetto, ma complesse come quelle di un essere umano, viene allora da una improvvisa propulsione degli studi nella robotica di nuova impostazione, anche incentivata da numerosi finanziamenti pubblici e privati (Brooks 2002). In sostanza, l'attimo andava colto proprio nel momento in cui era possibile farlo, senza rispettare pedissequamente il ritmo che l'evoluzione aveva impostato e che Brooks si era in prima istanza prefissato di seguire.

Il progetto di *Cog*, che beneficia non solamente della collaborazione di numerosi studiosi sotto la guida di Brooks, ma anche di filosofi di spessore come Daniel Dennett (1994, 1997), si identifica quindi come il primo vero esperimento di costruzione di un artefatto sintetico con le fattezze e i comportamenti di un essere umano *completo* (Brooks et al. 1998). *Cog* era composto solamente della parte superiore del corpo, che comprendeva testa, collo, braccia e parte del torso. Il cuore della sua architettura combinava l'impostazione reattiva già sviluppata nelle creature che abbiamo visto in precedenza con nuove tecniche di apprendimento derivanti dal ramo connessionista<sup>73</sup>

---

<sup>73</sup> Si è consapevoli di aver tralasciato lo sviluppo della corrente connessionista nella ricostruzione qui elaborata per una storia della robotica che sia informativa per un nuovo approccio di osservazione all'interazione tra umani e robot. La scelta è motivata da un focus sulle dinamiche che hanno portato alla comparsa degli approcci *behavior-based* e di robotica

dell'intelligenza artificiale: si voleva rendere capace Cog di apprendere e sviluppare in autonomia le proprie capacità di interazione con il mondo, che in questo momento per la prima volta comprende anche gli esseri umani che lo abitano e con cui è quindi possibile interagire. In particolare, Cog è considerato il primo esemplare di quella che sarebbe poi diventata l'approccio della *Developmental Robotics* (Lungarella et al. 2003; Cangelosi e Asada 2022), a volte definita anche *Epigenetic Robotics* (Prince e Gogate 2007), ovvero quel ramo della robotica che tenta di replicare i meccanismi di apprendimento tipici degli esseri umani al fine di indagarli e di costruire creature artificiali che ne dimostrino un'approssimazione (Lungarella et al. 2003: 151).

Il processo evolutivo che Brooks invocava su lunga scala con i suoi insetti si trasforma nel tentativo di riprodurre il processo di crescita e sviluppo che un essere umano affronta nel corso della sua vita, un passaggio dalla scala filogenetica a quella ontogenetica dell'evoluzione. In questa indagine, dunque, si concretizza un passaggio fondamentale agli occhi della nostra indagine: il processo di imitazione e simulazione dell'umano tramite la robotica arriva ad includere, grazie allo sviluppo della tecnologia a disposizione, anche i processi cognitivi sociali e la modellazione si concentra sulla conversione delle ipotesi elaborate in merito alla cognizione sociale nelle architetture cognitive dei robot. Nel caso di Cog, l'attenzione si concentra su quelle caratteristiche dell'apprendimento nel bambino che garantiscono l'acquisizione di competenze e abilità di interazione con il mondo e con gli altri sempre più complesse nel corso della sua vita. Un esempio è quello del tentativo di modellizzazione delle caratteristiche della *joint attention*, ovvero la capacità del bambino di rivolgere l'attenzione ad oggetti di interesse condiviso nel contesto dell'interazione sociale con il proprio caregiver (Meltzoff and Moore 1977; Tomasello e Farrar 1986). Ciò che si prospettava per Cog era la capacità di orientare il proprio "sguardo" verso il viso della persona con cui stava interagendo, di rintracciare degli stimoli salienti (come, ad esempio, un oggetto interposto tra il robot e l'umano) e mantenere l'attenzione verso l'oggetto di attenzione congiunta (Brooks et al. 1998: 71-72).

---

sociale, che si è creduto venissero meglio compresi se inquadrati nell'ottica di un'opposizione netta con l'approccio tradizionale della GOFAL. Tuttavia, è innegabile il contributo che l'approccio connessionista all'intelligenza artificiale abbia apportato allo sviluppo di una robotica situata e volta allo studio dell'interazione con gli umani. Per una disamina approfondita del rapporto tra connessionismo e robotica si rimanda a: Katič e Vukobratović 1996; Schlesinger 2009; Cangelosi 2017; Cangelosi e Asada 2022.



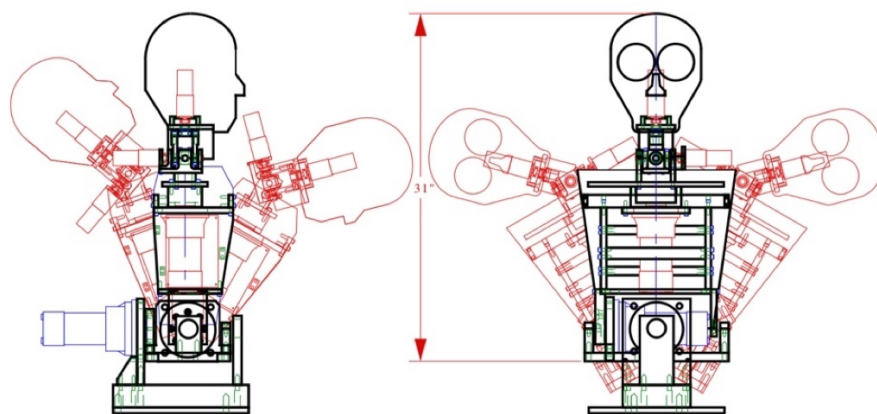


Fig 1.4 Alcuni dei movimenti di cui era capace Cog. Oltre a questi, Cog poteva anche muovere il collo a 180° e direzionare il torso a 120°. (Brooks et al. 1998: 62)



Fig. 1.5 Cog che rivolge la sua attenzione a un gesto della mano. Una volta che lo stimolo saliente è stato rintracciato, il robot rivolge lo sguardo all'oggetto e orienta la testa verso il target. (Brooks et al. 1998: 72)

In sostanza, il centro della ricerca portata avanti tramite Cog era lo studio dei meccanismi dello sviluppo inseriti all'interno di una prospettiva più ampia riguardo ai molti meccanismi di interazione sociale che permettono agli umani di interagire in modo naturale con altri esseri umani: ad esempio, Cog viene dotato della capacità di annuire se vede il suo interlocutore fare lo stesso (Scassellati 1998; Brooks et al. 1998: 78). La teoria di riferimento da cui il progetto estrae le ipotesi del funzionamento cognitivo da implementare è quella della *Theory of Mind* (ToM), in cui la cognizione sociale è vista come il meccanismo (a volte considerato innato) di attribuzione di pensieri, obiettivi, desideri, o altri stati intenzionali agli altri (Leslie 1994; Baron-Cohen 1995; cfr. §3.3). Il lavoro di operativizzazione, ovvero di conversione delle ipotesi concernenti il funzionamento di un processo cognitivo in una serie di proposte sul piano ingegneristico che siano capaci di implementare tali ipotesi, viene sviluppato per Cog da Brian Scassellati (1996, 2001, 2002, 2003) prendendo in considerazione meccanismi considerati fondamentali dalla ToM e rendendo possibile una qualità nell'interazione sociale mai raggiunta prima con dei robot: «A robotic system that possessed a theory of mind would allow for social interactions that have previously not been possible between man and machine» (2001: 13).

Per sviluppare ulteriormente lo studio dell'interazione sociale con robot umanoidi, Cynthia Breazeal decide di intraprendere la costruzione di un robot che, ancora meglio di Cog, potesse essere capace di realizzare interazioni simili a quelle che gli esseri umani hanno tra di loro. D'altronde, nel gruppo di Brooks si era già notato il fatto che, nonostante i loro robot dimostrassero ancora fattezze

molto meccaniche che permettevano senza dubbio di distinguere l'essenza robotica dei propri interlocutori e le loro abilità di interazione con il mondo non erano all'altezza di quelle umane, la maggior parte delle persone che si trovavano ad interagire con loro, o anche solamente alla loro presenza, tendevano a comportarsi in modi altamente sociali. Un famoso esempio è riportato dalla sociologa e psicologa Sherry Turkle, che incontra Cog per la prima volta:

Cog's mobile torso, neck, and head stand on a pedestal. Trained to track the largest moving object in its field (because this will usually be a human being), Cog "noticed" me soon after I entered its room. Its head turned to follow me and I was embarrassed to note that this made me happy. I found myself competing with another visitor for its attention. At one point, I felt sure that Cog's eyes had "caught" my own. My visit left me shaken – not by anything that Cog was able to accomplish but by my own reaction to "him". For years, whenever I had heard Rodney Brooks speak about his robotic "creatures", I had always been careful to mentally put quotation marks around the word. But now, with Cog, I had found that the quotation marks disappeared. Despite myself, and despite my continuing skepticism about this research project, I had behaved as though in the presence of another being. (Turkle 1995: 266).

L'indagine di Breazeal, allora, si concentra sui modi in cui il robot può suscitare un comportamento sociale da parte degli esseri umani, a volte anche sfruttando la cooperazione insita nel comportamento sociale umano<sup>74</sup>, e sviluppa un nuovo modello di robot che a Cog deve quasi tutta la sua architettura interna, ovvero Kismet (Breazeal 2002, 2003a, 2003b, 2004a, 2004b; Breazeal e Scassellati 2002). Kismet è un robot la cui sembianza ricorda quella di un giocattolo, per la precisione un *Furby*: non è un caso che i due oggetti si assomiglino, in quanto nello sviluppo del piccolo animaletto giocattolo era stato coinvolto anche il team di Brooks (Brooks 2002: 104-105). L'adozione della forma cartooresca del robot era motivata dal fatto che si potessero accentuare alcune delle caratteristiche facciali del robot, come gli occhi, le sopracciglia e la bocca, senza per questo sfociare nell'inquietante, oltre al fatto che i tratti accentuati avrebbero garantito una migliore specificazione delle emozioni espresse dal robot. Infatti, Kismet è stato uno dei primi esempi di *Affective Robotics*, che consiste nel tentativo di implementare l'espressione delle emozioni nei robot sociali, al fine di indagare le modalità affettive in interazione e comprenderne meglio la struttura (Kirby et al. 2010).

In sostanza, l'architettura di Kismet era costituita da sei sottosistemi che, seguendo l'impostazione delle architetture a sussunzione, portavano a compimento ognuno una competenza diversa (Breazeal 2004a: 44-46):

- *Low-Level Feature Extraction System*: questo sub-sistema è responsabile per il processamento dei dati provenienti dai sensori, che vengono trasformati in informazioni con una significatività per il comportamento del robot. A questo livello, sono modellizzati i comportamenti che

---

<sup>74</sup> Brooks racconta che l'idea di realizzare un robot che sfruttasse l'inclinazione sociale degli esseri umani sia venuta a Breazeal dopo essersi accorta che, nell'interazione con Cog, era lei a impostare l'interazione in modo che risultasse più sociale possibile, come nel caso della turnazione dell'azione (Brooks 2002: 92).

sembrano avere più importanza per lo sviluppo del bambino, come il rilevamento della direzione dello sguardo e il riconoscimento vocale.

- *Attention System*: a partire dai dati elaborati dallo strato inferiore, questo sistema seleziona tra quelli raccolti i dati salienti o particolarmente rilevanti in quel momento, al fine di rivolgere l'attenzione del robot verso questi e non altri stimoli. Il meccanismo di selezione dell'attenzione si basa su diversi criteri, come il fatto che uno stimolo appaia all'improvviso o si muova, oppure che sia considerato saliente di default (come lo sguardo). È in questo sistema che prendono forma le prime elaborazioni dei comportamenti emozionali del robot.
- *Perceptual System*: le informazioni fin qui elaborate dai sistemi inferiori vengono convogliate nel sistema percettivo, che sottopone i dati a una elaborazione in termini di significatività per il comportamento del robot. Il sistema ha dunque la funzione di accertarsi se le condizioni ambientali sono rispettate affinché possa essere attuato un certo comportamento.
- *Motivation System*: qui risiedono i criteri per stabilire gli "istinti" (*drives*) e le "emozioni" del robot. Gli "istinti" corrispondono ai "bisogni" che determinano lo stato del robot in termini di interazione: quando i bisogni sono soddisfatti, il robot si comporta seguendo le istruzioni dello stato in cui si trova (ad esempio, continuando a interagire con l'interlocutore); quando lo stato si allontana da quello desiderato, il sistema attiva delle procedure per riportare lo stato alla condizione desiderata, ad esempio attirando l'attenzione dell'interlocutore per incentivare l'interazione muovendo le orecchie o emettendo suoni. Le "emozioni" sono attivate sulla base di un dato stimolo e in comparazione con lo stato del sistema. Le possibilità di risposta sono sei (rabbia, disgusto, paura, gioia, dolore e sorpresa), che vengono messe in correlazione con risposte basate sulle soglie di attivazione date dallo stato (interesse, calma e noia).
- *Behavior System*: questo livello organizza i comportamenti del robot in una struttura coerente in base al task di riferimento. Ogni comportamento è visto come un'azione diretta a uno scopo e, dal momento che il robot può avere diversi obiettivi simultaneamente, un sistema di controllo regola l'attuazione del comportamento sulla base delle motivazioni elaborate nel livello inferiore. La responsabilità maggiore di questo sistema è quella di portare avanti questo meccanismo di decisione, considerando le caratteristiche in termini di rilevanza, coerenza, persistenza e opportunismo. Attraverso questo sistema, il robot è capace di comportarsi in maniera complessa in un ambiente dinamico.
- *Motor System*: Questo sistema garantisce al robot di dimostrare il comportamento attraverso movimenti ed espressioni. Consta a sua volta di quattro sottosistemi, che riguardano rispettivamente le abilità motorie, l'animazione facciale, la vocalizzazione espressiva e il comparto oculo-motorio. Quando sono selezionati un obiettivo e una strategia

comportamentale, il sistema motorio determina i movimenti posturali, la direzione dello sguardo, le emissioni vocali e le espressioni facciali.

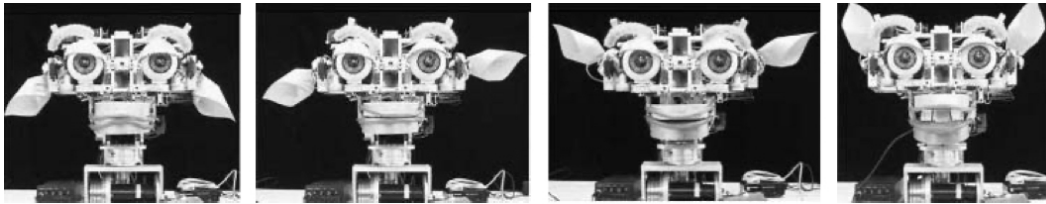


Fig. 1.6 Quattro delle espressioni di cui era capace Kismet. Da sinistra a destra: tristezza, disapprovazione, felicità, sorpresa. (Breazeal 2004a: 55)

Senza dubbio, la costruzione di robot come Cog e Kismet ha portato alla luce numerose questioni riguardo l'indagine dell'interazione tra umani e robot che fino a quel momento potevano ancora passare inosservate per via dello stato della tecnologia antecedente. Tra queste, si vuole sottolineare il doppio sguardo elaborato rispetto alla socialità nella robotica: da una parte, con Cog la ricerca tenta di implementare la capacità di cognizione sociale presentata di solito dagli umani in un robot che comprenda gli obiettivi dell'altro, sia capace di soddisfare le sue esigenze sociali e possieda capacità di "sintonizzazione" che forniscano feedback nell'interazione con l'umano; dall'altra, il lavoro di Kismet si concentra sulla sollecitazione tramite la manifestazione di emozioni delle capacità sociali dell'umano da parte del robot, il quale possa anche sfruttare dei processi "automatici" che avvengono nella interazione tra umani e robot, come l'antropomorfizzazione (cfr. §3.4). È a partire dai risultati raggiunti in questi progetti che il settore della robotica sociale si amplia fino a diventare un campo di studi autonomo tra gli altri all'interno dell'oramai variegato campo di indagine della robotica (Breazeal et al. 2016): l'aspetto della socialità dell'essere umano viene considerata non più una tra le tante capacità dimostrate da un sistema intelligente, ma uno degli aspetti cardine per la costruzione di robot che dimostrino un'intelligenza e un comportamento complessi.

Arrivati fin qui, sembra che il percorso intrapreso dalla Nouvelle AI sia vincente su tutti i fronti: le creature sviluppate da Brooks e colleghi sembrano funzionare e comportarsi meglio di tutti i loro predecessori, riescono a innescare interazioni *human-like* quando interagiscono con gli esseri umani, e sembrano confermare l'ipotesi che i principi teorici descritti all'inizio di questo paragrafo fossero sia necessari sia sufficienti per realizzare robot che dimostrino un comportamento intelligente. Ma qual è il loro ruolo all'interno del processo sintetico di verifica delle ipotesi di partenza?

### 2.3.3 Modelli del mondo e modelli di mente

È raro trovare un testo che, se affronta anche solamente parte della storia legata alla costruzione moderna di artefatti artificiali, non tenga conto del contributo di Brooks e della sua squadra di

collaboratori nel campo della robotica. È innegabile, come abbiamo provato a mostrare, che l'operazione svolta da Brooks abbia contribuito enormemente a sviluppare lo stato dell'arte della robotica come disciplina ingegneristica di creazione di artefatti artificiali somiglianti all'umano, abbandonando i dogmi propri della GOFAI e snellendo la procedura di progettazione situandola nel mondo fisico. Tuttavia, ciò che ancora meglio restituisce la matrice rivoluzionaria del contributo di Brooks è la sua comparsa in molti testi riguardanti la storia delle scienze cognitive. Infatti, il successo avuto dalla Nouvelle AI ha travalicato i limiti della disciplina in cui era nata e ha informato le riflessioni filosofiche in merito allo scenario teorico ed epistemologico dei nuovi approcci alle scienze cognitive, rafforzando la corrente alternativa che vedeva la cognizione a partire dalla sua origine biologica ed esperienziale.

Il primo esempio di attestazione del contributo alle scienze cognitive dell'opera di Brooks è rintracciabile nell'opera fondatrice di una corrente radicale delle scienze cognitive, che affonda le sue radici nella teoria dell'autopoiesi di Maturana e Varela, e che ha tra i suoi fondatori uno dei suoi ideatori. Infatti, nel volume *The Embodied Mind* (Varela et al. 1991), considerato il testo cardine della teoria enattivista della cognizione, l'impostazione di Brooks per la robotica è annoverata tra i primi esempi (*ante litteram*) della corrente filosofica che in quel testo veniva fondata: l'*enattivismo*. I due approcci condividono molti, se non tutti, i presupposti epistemologici alla base dell'indagine della mente e della cognizione: anti-rappresentazionalismo radicale, enfasi sul ruolo del corpo e dell'esperienza nella costituzione delle capacità cognitive dell'organismo, rifiuto della metafora computazionalista, riferimento primario alla biologia come dominio disciplinare a cui afferire per considerare la cognizione un fenomeno naturale, stretto legame – portato fino all'identità dei termini per la corrente enattivista (Ibid.: 238-239) – tra percezione e azione. Presentando le ricerche riguardo gli insetti di Brooks, Varela e colleghi mostrano come i presupposti alla base della ricerca sulle creature robotiche fino ad allora costruite (nel 1991 non erano ancora in programma né Cog, né Kismet) fossero da prendere in considerazione come i risultati più promettenti nel campo dell'intelligenza artificiale e in quanto attestazione della validità delle ipotesi alternative al pensiero computazionale e simbolico.

Questo esempio di ciò che definiamo intelligenza artificiale enattiva è distintamente e chiaramente formulato come tale dai suoi sostenitori (che naturalmente non usano il nostro termine 'enattivo'). Come afferma lo stesso Brooks, il suo orientamento non è né una forma di connessionismo, né di produzione di regole, né di ermeneutica. È motivato dagli stessi cari vecchi interessi ingegneristici che hanno prodotto il cognitivismo e il connessionismo. Sono proprio questi interessi ingegneristici a rivelare più chiaramente come il concetto di cognizione come enazione stia scaturendo dalla stessa logica di ricerca e di sviluppo nelle attuali scienze cognitive. L'approccio enattivo, allora, non è una preferenza meramente filosofica, bensì il risultato di forze interne alla ricerca nelle scienze cognitive, anche nel caso di quegli ingegneri pragmatici che vogliono costruire macchine davvero intelligenti e utili. (Varela et al. 1991: 283)

Ci troviamo a concordare con Varela, Thompson e Rosch in merito all'utilità delle creature sviluppate grazie all'approccio *behavior-based*: questo genere di architetture ha reso effettivamente più robusti i comportamenti esibiti dai robot nel mondo che abitiamo, permettendo a queste macchine di muoversi molto più agilmente dei loro predecessori anche in ambienti dinamici e non controllati, molto più simili al nostro. Ciò che invece non può essere affermato con tanta tranquillità è che questo approccio abbia condotto alla costruzione di creature «davvero intelligenti». I progetti portati avanti da Brooks, infatti, dopo una impressionante notorietà iniziale, si arenano negli anni successivi fino a diventare, sia gli insetti sia Cog sia Kismet, delle opere da museo (Boden 2006: 1303-1304). Soprattutto per quanto riguarda Cog, le aspettative iniziali riguardo la possibilità di raggiungere un livello di intelligenza umanoide hanno dovuto fare i conti con il fatto che la complicazione e la moltiplicazione dei livelli di comportamento non fosse una via diretta per il raggiungimento della complessità globale del sistema: l'idea che l'intelligenza fosse distribuita incrementalmente in linea di continuità secondo una direttrice che andava dagli insetti agli esseri umani aveva tratto in inganno sia Brooks che Dennett (Dreyfus e Dreyfus 2008), conducendo a rilento i miglioramenti nel periodo di sviluppo e portando poi all'abbandono del programma.

Proprio a partire da tali constatazioni, è utile considerare alcune riflessioni provenienti dall'approccio filosofico che ha informato la stessa costruzione di questi artefatti in merito ai problemi per cui l'approccio di Brooks si credeva avesse trovato delle soluzioni. In particolare, il discorso si rivolge di nuovo alla questione del *frame problem*. Nell'argomentazione di Brooks, il *frame problem* era concettualizzato come l'impossibilità dei sistemi artificiali simbolici di ancorare la propria attività alle condizioni reali del mondo: come abbiamo visto, questo era causato dal fatto che il mondo in cui questi sistemi navigavano era un mondo giocattolo, costruito a posta per loro e che non aveva corrispondenza diretta nelle condizioni del mondo fisico. Quindi, attraverso la proposta di una robotica ancorata fisicamente, il *frame problem* sembra poter svanire, dal momento che era stato creato proprio dalle condizioni messe in opera dagli account simbolici. In aggiunta, anche altri problemi che si erano venuti a creare all'interno del paradigma computazionalista simbolico sembravano potersi dissolvere, come ad esempio il *symbol grounding problem* (Harnad 1990), che riguarda il modo attraverso cui i simboli acquisiscono il loro significato: se non ci sono simboli e l'interazione è situata all'interno dell'ambiente, non c'è ragione di porre il problema. Come sottolineato da Pfeifer, nell'interpretazione della nuova corrente dell'AI la creazione di tali questioni era riconducibile alla scarsa comprensione della natura dei sistemi intelligenti (1995: 52), approfondita la quale i problemi sarebbero svaniti da soli.

Contrariamente a quanto prospettato, però, la questione si è rivelata più complessa. Non solo le critiche rispetto a questa soluzione sono arrivate dal versante della scienza cognitiva classica, i cui

esponenti vedevano in questo tipo di architetture solamente la possibilità di ottimizzare l'implementazione di meccanismi cognitivi di livello inferiore (come la navigazione e la reazione agli stimoli) ma non quelli di livello superiore (come il linguaggio e il ragionamento astratto) (Pylyshyn 1987; Ford e Pylyshyn 1996), ma anche dagli stessi che avevano contribuito a informare il nuovo approccio alle scienze cognitive. In particolare, Dreyfus individua un punto debole nella argomentazione di Brooks riguardo alla soluzione trovata per il *frame problem*.

Brook's approach is an important advance, but Brooks' robots respond only to fixed isolable features of the environment, not to context or changing significance. [...] Brooks thinks he does not need to worry about learning, putting it off as a concern for possible future research. But by operating in a fixed world and responding only to the small set of possibly relevant features that their receptors can pick up, Brooks' [robots] beg the question of changing relevance and so finesse rather than solve the frame problem. (Dreyfus e Dreyfus 2008: 336).

Il problema veniva riportato da Dreyfus ad un approccio fenomenologico, in cui l'esperienza, e non solo il fatto di avere un corpo situato nell'ambiente, garantisce la comprensione della situazione e la possibilità di agire efficacemente al suo interno senza che ci sia la necessità di un meccanismo di pianificazione o controllo centralizzato. In fondo, fa notare Dreyfus, già Brooks aveva paventato la possibilità che il suo approccio fosse parziale rispetto ai problemi sorti fino a quel momento nel panorama dell'intelligenza artificiale, non perché si prospettasse sbagliato, ma data la stessa natura tentativa dell'impresa umana di conoscenza:

Perhaps we have all missed some organizing principle of biological systems, or some general truth about them. Perhaps there is a way of looking at biological systems which will illuminate an inherent necessity in some aspect of the interactions of their parts that is completely missing from our artificial systems. This could be interpreted to be an *elixir of life* or some such, but I am not suggesting that we need go outside the current realms of mathematics, physics, chemistry, or bio-chemistry. Rather I am suggesting that perhaps at this point we simply do not *get it*, and that there is some fundamental change necessary in our thinking in order that we might build artificial systems that have the levels of intelligence, emotional interactions, long term stability and autonomy, and general robustness that we might expect of biological systems. In deference to the elixir metaphor, I prefer to think that perhaps we are currently missing the *juice* of life. (Brooks 1997: 301)

Nel delineare la critica all'approccio di Brooks, allora, Dreyfus individua una sorta di timidezza nel modo in cui sono stati applicati in concetti provenienti dall'approccio alternativo delle scienze cognitive nella pratica ingegneristica di creazione delle creature artificiali: l'impresa di una intelligenza artificiale heideggeriana ha sì fallito nello spiegare tutto quello che si era prefissata fin dal principio, ma solamente perché era troppo poco heideggeriana (Dreyfus e Dreyfus 2008). In ogni caso, il *frame problem* sembra non essere stato risolto nel passaggio da GOFAI a Nouvelle AI, e anzi sembra essere ancora molto pertinente sia nell'indagine sull'intelligenza artificiale (robotica compresa) (Froese e Taguchi 2019), sia nelle scienze cognitive (Froese e Ziemke 2009).

A fronte di queste critiche, che mettevano in risalto alcuni punti deboli del tentativo di rispondere a problemi già emersi nel campo dell'intelligenza artificiale e della robotica, c'è da notare inoltre l'apertura di un dibattito relativo alla capacità dei robot di poter implementare i principi assunti da Brooks alla base della sua modalità di costruzione. L'obiettivo di tale linea di ricerca si sviluppa cercando di analizzare in profondità le implicazioni dei nuovi concetti introdotti dalla *Nouvelle AI*, come ad esempio la possibilità di escludere totalmente l'uso di rappresentazioni nella costruzione delle architetture (Steels 2003) o la concettualizzazione dell'autonomia del comportamento robotico raffrontata alle definizioni date dalla letteratura sull'autopoiesi (Arkin 1998; Smithers 1997). Tra questi, il principio approfondito più nel dettaglio è quello relativo all'*embodiment* dell'organismo: la questione se un robot possa essere effettivamente considerato come un agente incarnato e situato nel proprio mondo-ambiente è affrontata confrontando le forme di comportamento e di elaborazione tra diversi tipi di organismi. Tale riflessione condurrà in seguito alla necessità di formulare una gerarchia dei gradi di *embodiment* che gli agenti possono dimostrare di avere in relazione al proprio ambiente, relegando i robot costruiti da Brooks a incorporare solamente stati basilari e "deboli" del processo (Sharkey & Ziemke 1998, 2000; Ziemke 2003a, 2003b, 2004, 2007, 2016, 2022; Ziemke & Thill 2014, §3.4.2).

Ora, è proprio a partire dall'osservazione del modo in cui lo sviluppo del paradigma di Brooks è stato accolto all'interno delle scienze cognitive e della filosofia a queste dedicate che possiamo affrontare l'ultimo obiettivo di questo paragrafo, ovvero quello di guardare al ritorno delle applicazioni robotiche sulla teoria che le ha informate come movimento induttivo del metodo sintetico. Possiamo analizzare il nostro viaggio di ritorno – dalla robotica alle scienze cognitive – attraverso una serie di tappe intermedie:

- In un primo momento, l'elaborazione delle critiche al paradigma della GOF AI si raccoglie intorno a figure cardine come quelle di Weizenbaum, Winograd e Dreyfus, i quali sviluppano un approccio alternativo i cui presupposti non possono essere più diversi da quelli delle scienze cognitive computazionaliste e simboliche. Tale insieme di critiche, però, rimane secondario nel campo dell'Intelligenza Artificiale e della robotica, che continuano a sviluppare il loro approccio attraverso concetti come quelli di pianificazione e rappresentazione centrali, considerando la controcultura come un singulto congenito di matrice fenomenologica.
- Allo stesso tempo, il recupero di alcune idee provenienti dalla cibernetica da parte del BCL e correnti di pensiero affini consente di riallacciare i rapporti tra una concezione teorica della cognizione basata sulla biologia e l'impresa di costruire artefatti artificiali che dimostrino un'intelligenza umana. L'introduzione nel dibattito ufficiale dell'AI di nozioni come quelle di auto-organizzazione, autopoiesi e autonomia da parte di Winograd (e Flores) conducono a



riconsiderare la possibilità di discostarsi dagli approcci tradizionali alle scienze della mente e dell'artificiale non solo sul piano teorico, come faceva Dreyfus senza proporre soluzioni alternative di natura pratica, ma anche su quello ingegneristico.

- È così che viene resa possibile l'emersione dell'approccio *behavior-based* elaborato da Brooks, che si nutre delle formulazioni alternative agli approcci classici sviluppate fino a quel momento e rilancia la sfida radicalizzando le idee e realizzando artefatti che veniva dichiarato funzionassero proprio secondo queste ipotesi. Il successo dimostrato dalle creature di Brooks si riscontra non solo nella robotica, che non avrebbe fatto più a meno delle concettualizzazioni della *behavior-based robotics* (Pfeifer & Scheier 2001; Pfeifer e Bongard 2006; Mataric 2007), ma anche nelle scienze cognitive, che fin all'inizio degli anni '90 inaugurano una completa ristrutturazione mettendo al centro le nozioni di corpo, percezione e azione con l'*embodied turn* (Varela et al. 1991; Clark 1998; 1999). Le critiche elaborate a seguito delle problematiche riscontrate da Brooks vengono fatte derivare non dall'impostazione teorica ed epistemologica, ma dalle difficoltà di operativizzazione e implementazione dei principi per la costruzione di un artefatto artificiale.
- La considerazione delle creature di Brooks all'interno delle scienze cognitive porta non solo all'arricchimento del panorama teorico, ora non più concentrato solamente sugli aspetti logici e disincarnati della cognizione, ma anche ad un affinamento delle categorie elaborate dalla corrente alternativa alla GOFAI e a delle considerazioni critiche attraverso cui si rende possibile una ulteriore specificazione delle idee e della loro possibile operativizzazione. Ad esempio, alcune riflessioni sul concetto di *embodiment* a partire dalla loro presunta realizzazione negli artefatti di Brooks conducono a una segmentazione dei possibili significati del termine e a una gerarchizzazione delle possibilità di situazione incorporata degli agenti – biologici e artificiali – nel mondo.

In sostanza, ciò che viene preso maggiormente in considerazione dalle scienze cognitive è l'aspetto teorico, supportato da alcune realizzazioni tecnologiche, dell'impresa di Brooks. Infatti, ancor più importante del fatto che le sue creature si mostrassero effettivamente capaci di muoversi agilmente nel mondo, l'enfasi posta sulla variazione epistemologica che ne aveva reso la realizzazione attuabile ha contribuito a:

- i) rinsaldare le posizioni in merito all'importanza del corpo e dell'esperienza;
- ii) considerare l'intelligenza da prospettive diverse rispetto a quella antropocentrica, ora analizzata in base al contesto di azioni di riferimento, abbandonando la pretesa di generalità e concentrandosi invece sulle soluzioni trovate nel contesto pragmatico;

- iii) concepire la cognizione come una azione strettamente legata alla percezione;
- iv) riconsiderare il ruolo del comportamento per l'interazione con gli artefatti artificiali.

Inoltre, e non tangenzialmente per il nostro discorso, l'emersione di nuove forme di indagine legate alla natura sociale degli artefatti artificiali resa possibile da questi sviluppi tecnologici ha inaugurato un dominio di ricerche nelle scienze cognitive che oggi è più florido che mai e che riguarda il modo in cui gli esseri umani interagiscono con i robot (Dautenhahn 2007; Kanda & Ishiguro 2013; Bartneck et al. 2019). In sostanza, le scienze cognitive hanno assunto le creature di Brooks, dagli insetti a Kismet, come una prova della validità delle ipotesi che incorporavano, inglobando al loro interno i presupposti che erano stati dimostrati essere utili per la costruzione di artefatti capaci di mostrare intelligenza. Se si vuole riassumere in una frase la tendenza delle scienze cognitive a trovare rinforzo nelle proprie tesi tramite realizzazione di artefatti artificiale, si può dire, come fatto da Boden (2006: 176), che, piuttosto che ancorarsi a idee statiche per descrivere la mente, «It's more accurate to say [...] that cognitive science aims to describe the mind by *concepts drawn from the best theory of what computers do – whatever that theory turns out to be*».

In questo viaggio, quello che vogliamo portare a casa come souvenir è una riformulazione del modo di vedere la robotica in quanto impresa ingegneristica di realizzazione dell'ambizione dell'umano: non è solo considerando le caratteristiche astratte del pensiero e della mente che è possibile informare la costruzione di robot che agiscano efficacemente e in modo intelligente nel mondo e in interazione con gli umani; è necessario includere anche fattori che a prima vista possono sembrare superflui, ma che si rivelano cruciali per garantire un'efficacia pragmatica.

Conclusasi qui la nostra traversata tra scienze cognitive e robotica, possiamo formulare le tesi che ci accompagneranno nei prossimi capitoli a partire da alcune osservazioni che è possibile sviluppare riguardo la robotica sociale e il rapporto con gli esseri umani.

## **2.4 Nuove strade da intraprendere: filosofia della complessità e *Human-Robot Interaction***

Per comprendere al meglio la direzione che vogliamo intraprendere, è utile concludere questo capitolo ripartendo da dove abbiamo iniziato. In *Westworld*, alla domanda posta dal robot Bernard «Cosa mi rende diverso da te?», il suo costruttore gli risponde che, semplicemente, «non c'è niente che ci rende più della somma delle nostre parti»: siamo tutti macchine, scoperto il funzionamento delle quali è possibile istituire un piano di uguaglianza assoluto su cui confrontarci rispetto alle somiglianze, più che alle differenze. In questa narrazione, l'ambizione dell'umano di comprendere se stesso attraverso il proprio riflesso artificiale trova pieno compimento grazie al potere del racconto e della finzione. Bernard realizza, davanti agli occhi dello spettatore, il sogno antico della replicabilità

e quello moderno di una comprensione dell'essenza umana tramite la costruzione della macchina; il creatore Ford, nel suo ruolo, incarna e rinnova i tratti di figure archetipiche come Prometeo, Efesto, e il dottor Frankenstein: tutti esseri che, nella mitologia o nella letteratura, hanno cercato di comprendere l'essenza umana tramite la produzione di artefatti capaci di riflettere il loro stesso creatore.

A seguito del percorso che abbiamo compiuto, possiamo riconoscere che nell'affermazione di Ford pulsa nel profondo la lezione della metafora computazionale. *Westworld* è ambientato in un mondo, futuro o parallelo, in cui la metafora della mente come particolare tipo di macchina – il computer – si è rivelata la giusta via da perseguire per replicare la natura della coscienza, la capacità emotiva e l'intelligenza dell'essere umano: un mondo in cui niente è più della somma delle sue parti.

Tuttavia, come sappiamo, questo mondo sembra ancora di là da venire. Nel nostro mondo, in cui la versione attuale della metafora computazionale non ha – forse, solo per adesso – portato alla realizzazione di creature artificiali coscienti e capaci di esperienza, non abbiamo corrispettivi artificiali che possano domandarci cosa li rende diversi da noi, non abbiamo Bernard che ci interrogano direttamente sul margine che ancora resta da appianare tra umano e macchina. E, tuttavia, il confronto con le creature artificiali alimenta comunque l'ambizione dell'umano di comprendersi attraverso la macchina, come abbiamo visto nell'operazione di costituzione delle scienze cognitive e della robotica, anche se queste sembrano non poter rispondere nel modo in cui farebbero artefatti come quelli in *Westworld*. Dobbiamo forse arrenderci al fatto che, per riuscire a indagare le nostre abilità e la nostra mente attraverso la macchina, si debba attendere che il proposito della metafora computazionale si realizzi? Dobbiamo necessariamente completare l'operazione di replicazione per far sì che l'indagine della robotica acquisti un suo senso specifico?

Il parallelo tra mondo della narrativa e realtà tecnologica suggerisce un'altra modalità di interrogare il robot, che non ha bisogno di livelli di coscienza superiore come quelli sviluppati da Bernard per essere messa a frutto: nel dialogo tra costruttori e creature artificiali, quello che possiamo interrogare è proprio il momento in cui forme diverse di intelligenza e cognizione si incontrano. Ecco, allora, che possiamo cogliere uno spunto da questa narrazione, non perseguendo l'indagine della mente e della cognizione sotto la guida della metafora computazionale, ma indirizzando il percorso di indagine verso nuove possibilità di scoperta: comprendere la struttura del dialogo tra umani e robot.

Il percorso che abbiamo affrontato ci permette di ribaltare la risposta di Ford e rivalutare *Westworld* sotto la prospettiva dell'interazione tra umani e robot. Sebbene, infatti, il focus di tutta la serie ruoti attorno ai temi del doppio, dell'imitazione e del pericolo che deriva dalla volontà di ricoprire il ruolo di Prometeo contemporanei, l'attenzione può essere spostata sull'impiego che gli esseri umani fanno dei robot. Bernard e Ford istituiscono un dialogo in cui è messa in risalto la possibilità di una

reciprocità dell'indagine: da una parte, l'umano indaga il suo artefatto, il quale riesce a dire qualcosa sul livello di comprensione raggiunta rispetto a sé stesso; dall'altra, il robot indaga l'umano, per rilanciare domande circa le differenze che ancora li separano. È in questo dialogo interazionale che i robot, anche se molto diversi da noi, possono darci conto della qualità della nostra comprensione di noi stessi, restituendoci sotto un preciso rispetto il prodotto delle nostre ipotesi messe alla prova; è nello stesso dialogo che ritroviamo la possibilità di indagarci, riscoprendo continuamente differenze laddove cerchiamo somiglianze.

Il nuovo percorso che vogliamo aprire, dunque, si configura come il tentativo di raccogliere una *sfida della complessità* (Bocchi e Ceruti 1985; Morin 2011), che tenga in considerazione la questione dell'interazione tra umani e robot su diversi livelli, a partire da una scienza cognitiva che mette in relazione il contributo determinante della corporeità, il ruolo dell'esperienza e la fondamentale relazione tra percezione e azione, al fine di comprendere meglio come sia possibile costruire creature artificiali che possano rispettare le nostre aspettative e collaborare in accordo al modo in cui noi ci predisponiamo. In questo contesto, l'iscrizione dell'indagine all'interno del dominio della filosofia delle scienze cognitive enattiviste contemporanee (Varela et al. 1991; Damiano 2010; Thompson 2007; Di Paolo et al. 2010; Hutto & Myin 2013, 2017) consente di informare la costruzione di artefatti robotici non solamente sul paragone con l'interazione fra umani, ma sviluppando uno sguardo che tenga presente, in primo luogo, l'unicità dell'interazione in oggetto. L'impresa di creare robot che siano dei perfetti doppi degli esseri umani, tentando di incorporare l'ultima ipotesi sviluppata nel campo delle neuroscienze o delle scienze cognitive, è un'operazione che affonda le proprie radici nella metafora della cognizione come computer in cui l'essere umano – con la sua intelligenza – è una macchina computazionale, trovato il segreto della quale può essere riprodotta su qualsiasi altro supporto che ne garantisca l'implementazione. Il percorso che intendiamo seguire, d'altra parte, colloca il tentativo di costruire esatti doppi robotici in un quadro più ampio di analisi della natura e della cognizione umane, che risulta sempre parziale rispetto alla generalità del modo in cui l'essere umano comprende, agisce, dà senso al mondo. Ciò che ci proponiamo di fare è invece istruire l'impresa robotica con uno sguardo che consenta di porre al centro l'essere umano e il tentativo di comprensione della sua natura interazionale, specificata nelle sue componenti esperienziali, corporee e pragmatiche.

Il percorso che vogliamo intraprendere, allora, parte dal presupposto che se vogliamo ottimizzare il modo in cui i robot, in quanto forme di imitazione delle nostre stesse capacità di comprensione e comunicazione, si avvicinano ad interagire con noi, è necessario guardare al modo in cui noi – in quanto esseri umani – interagiamo con loro. Il viaggio che abbiamo appena compiuto si configura come una lezione di cui possiamo fare tesoro per intraprendere questo nuovo percorso, che ci conduca

a una comprensione più profonda del rapporto che lega gli esseri umani alle loro forme di imitazione/simulazione robotiche. In questo viaggio, dunque, il procedimento di costruzione della comprensione dovrà tenere presente alcuni presupposti:

- i) La metafora computazionale ancora oggi informa una corposa corrente delle scienze cognitive, modificata nel corso del tempo dalla comparsa di nuove forme di computazione, nuove tecnologie e possibilità. In questa congerie, il rapporto che lega l'essere umano alle metafore che usa per comprendere sé stesso è strettamente dipendente dallo sviluppo delle tecnologie che ha a disposizione nel proprio scenario interpretativo. Tuttavia, la nostra ricostruzione ci ha consentito di vedere come sia opportuno considerare l'essere umano non (necessariamente) attraverso metafore di natura computazionale, ma affiancarle o sostituirle con analogie che provengono dal dominio biologico, che vede l'essere umano come sistema auto-organizzato e capace di dare senso al mondo nel processo di mantenimento della propria autonomia e dei suoi stati di dis-equilibrio. Il contributo della filosofia dell'autopoiesi ci permette di immaginare l'essere umano non più come una macchina deterministica, da scomporre e ricomporre per scoprire il segreto del suo funzionamento, ma come un complesso capace di costruire continuamente le sue condizioni di possibilità in costante accoppiamento con il mondo che lo accoglie e lo modella. La metafora che vogliamo adottare, allora, è quella che si costruisce sempre localmente attraverso la composizione delle caratteristiche proprie di un sistema, delle sue capacità interazionali e della situazione in cui il sistema stesso si trova ad agire. Come sottolineato da Israel (2004: 69-70),

[...] i risultati più importanti nella storia della concezione dell'uomo-macchina sono sempre consistiti nella realizzazione di macchine utili ed efficaci, più che nella dimostrazione che l'uomo è una macchina, rispetto alla quale ci troviamo sempre allo stesso punto.

Nella ricerca condotta sotto il segno della metafora computazionale, non potremo che trovare i meccanismi che accomunano computer e intelligenza umana. Seguiremo, allora, una strada alternativa, che non provi a ridurre l'umano a una macchina ma cerchi di analizzare il modo in cui gli umani come esseri viventi, macchine autopoietiche, incontrano robot, macchine computazionali, e si trovano a intraprendere percorsi di costruzione del senso congiunto sulla base di modalità tutte umane di comprensione.

- ii) Il passaggio da una prospettiva della generalità, della misurabilità, del definito, a una del tentativo, dell'eccedenza, del contingente, è necessario per inquadrare l'indagine sotto il rispetto della particolarità dell'interazione tra umani e robot, non cercando di trovare uniformità laddove sono più salienti le differenze con l'interazione che avviene tra umani,

e viceversa. Perciò, il contributo dell'esperienza e il ruolo dell'osservatore devono essere costantemente tenuti in considerazione, cercando di orientare l'indagine verso una specificità delle caratteristiche salienti invece che verso una generalità delle regole del gioco. La considerazione del ruolo dell'osservatore nell'impresa di costruzione di artefatti artificiali, come sottolineato da von Foerster esplicitamente, ci ha condotto ad apprezzare il tentativo di mettersi nella posizione di chi guarda, cercando allo stesso tempo di considerare il sistema osservante come parte del sistema oggetto di indagine, e non come uno sguardo disincarnato che analizza la situazione in terza persona. L'indagine dell'interazione tra umani e robot che proponiamo terrà in considerazione il modo in cui le teorie della cognizione sociale hanno compreso la natura interazionale dell'essere umano, ma cercherà di declinarla nella specificità dell'incontro con i robot. Il modo in cui si presenta il comportamento in interazione, allora, costituisce il nostro oggetto di indagine, secondo una duplice presa in considerazione: da una parte, guarderemo a come la nostra esperienza di fronte al robot sia condizionata e informata da comportamenti e abitudini sviluppate nel corso della nostra esperienza sociale con altri esseri umani; dall'altra, cercheremo di declinare la specificità dell'interazione con il robot a partire dalle considerazioni delle sue sembianze e del suo costituirsi come un agente sociale.

Nel quadro che abbiamo appena delineato, l'indagine dell'interazione tra umani e robot si configura come l'occasione di ripensare alle teorie della mente e della cognizione prendendo in considerazione come diverse forme di intelligenza, artificiale e biologica, possano istituire piani di incontro e comunicazione senza necessariamente condividere le stesse modalità cognitive. La metafora computazionale costituisce ancora oggi il paradigma maggioritario delle scienze cognitive, e instrada la costruzione di artefatti robotici guidata dall'ipotesi che la modellizzazione computazionalista delle funzioni cognitive possa, in qualche modo, fornire il terreno comune del funzionamento di entrambe le forme di intelligenza. Affinché possiamo comprendere come l'umano ha impostato lo studio dell'interazione con i robot sotto questa corrente, nella prossima sezione forniremo una disamina di alcune delle tematiche al centro degli studi di *Human-Robot Interaction* (HRI) e cercheremo di costruire alternative all'interpretazione maggioritaria mostrando come le teorie radicali della cognizione *embodied* possano aprire nuove strade e concepire l'interazione come un terreno di incontro tra diverse forme di comprensione. Come vedremo, queste teorie offrono una prospettiva secondo cui la cognizione non risiede più esclusivamente all'interno della mente individuale, ma si manifesta nell'interazione con l'ambiente e con gli altri agenti che lo abitano, dove processi di senso emergono grazie a un coinvolgimento attivo ed *embodied* delle capacità cognitive

dell'agente. Questo cambiamento di prospettive favorirà la riformulazione di molte posizioni all'interno degli studi sulla HRI, che includeranno sia approcci alternativi allo sguardo generale, sia riconsiderazioni locali di tematiche specifiche.

## **Parte II -**

### ***Ti estì.* Il robot agli occhi dell'osservatore**



## Capitolo III – Galatea Moderna.

### Robot sociali tra modellizzazione e imitazione del vivente

[...] Pigmalione  
ne rimane incantato e nel suo cuore  
brucia d'amore per quel corpo finto.  
Spesso passa la mano sulla statua  
per constatare se davvero è carne  
e non avorio (ma non vuole ammettere  
che sia d'avorio), la bacia, le parla,  
e s'illude che lei lo corrisponda,  
e quando poi l'abbraccia ha l'impressione  
che le sue dita affondino nel corpo  
che va toccando e teme che lo stringerla  
possa lasciare lividi sugli arti.  
Ovidio, *Metamorfosi*, Libro X (vv. 253-258)

#### 3.1 Epistemologia di una nuova interazione

Il mito di Pigmalione è per la prima volta riportato da Ovidio nel suo celebre volume sulle *Metamorfosi*. Nel libro X, il poeta narra di uno scultore cretese che, considerando le donne del suo paese come degenerazioni della natura umana, si ritira celibe alle sue attività artistiche e passa la maggior parte della sua vita isolato nel suo laboratorio a scolpire statue dalla forma umana. Ciò accade fino a che lo scultore non realizza una statua in avorio di fattura così pregiata e con fattezze talmente realistiche «quam vivere credas» (Ov. *Met.* vv. 250). Pigmalione se ne innamora perdutamente, entrando in uno stato di adorazione della sua stessa opera. L'epilogo della storia è forse ancora più celebre del suo inizio: pregando Afrodite che lo accontenti nel desiderio di amare la sua creazione in carne ed ossa, la dea dà vita alla statua, che si presenta come una donna con la pelle diafana e le forme perfette, e che dà in sposa al suo creatore.

Fin dalle sue prime formulazioni, e con riprese diffuse in tutta la letteratura occidentale (Hersey 2009), il mito ha contribuito a definire i canoni del rapporto tra creatore e opera d'arte, alimentando l'ambizione dell'umano di accedere al segreto della vita, prerogativa divina, e a ritagliare le pertinenze attraverso cui leggere il rapporto tra chi crea e l'oggetto creato. Infatti, come abbiamo visto, il mito di Pigmalione è uno di quei racconti che ha contribuito a nutrire un immaginario in cui la tecnica nelle mani dell'essere umano porta in qualche modo alla realizzazione di altri esseri umani: in questo caso specifico, lo scultore è meritevole dell'intervento divino data l'estrema qualità della sua opera e l'uso pregevole del suo talento scultorio.

È interessante notare che, fino a tempi moderni (Law 1932), alla statua non è mai stato dato un nome. La versione di Ovidio non ne fa parola e la letteratura si è sempre riferita alla creazione di Pigmalione in termini vaghi, concentrando l'attenzione sulla figura dello scultore e sul suo ruolo di ammaestratore della materia tramite la tecnica. Solamente con la ripresa del mito in epoca romantica, la statua acquista il nome di Galatea. Da quel momento in poi, forse proprio in seguito ad uno spostamento dell'attenzione verso l'artefatto, la statua acquista una sua identità, una sua volontà e una sua dignità nel racconto del mito. Evidenze moderne di questo processo sono le sue rielaborazioni contemporanee (Essaka 2001; Miller 2013; Flamm 2019), che mettono sempre più al centro la figura della statua e le attribuiscono un ruolo centrale rovesciando i rapporti di potere. Galatea non è più solamente il simbolo della tecnica che porta alla vita grazie all'intervento magico a volte divino, non esaurisce il suo significato nell'essere obiettivo ultimo delle capacità dell'umano – ricreare la vita – ma si distingue nel racconto con la sua capacità di interfacciarsi, di comportarsi in modi coerenti alle situazioni, di dimostrare un'autonomia significativa delle azioni che fa la differenza tra la vita e una sua riproduzione. Il mito non esaurisce più il suo potere nella strutturazione del rapporto tra creatore e opera, ma si espande arrivando a comprendere le caratteristiche che Galatea deve incarnare in quanto nuovo umano: sapere interagire.

È in questo punto che, allora, possiamo riallacciare il discorso che abbiamo iniziato nella parte precedente e cogliere un suggerimento riguardo a come inquadrare l'analisi dell'interazione tra umani e robot attraverso lo sguardo dato dal mito di Galatea. In principio, Pigmalione è l'unico protagonista di questa storia, il solo vero agente all'interno dello svolgersi degli eventi, il cui desiderio è il motore che consente alla statua di acquistare una "umanità". Nell'evoluzione del mito, tuttavia, la statua si autonomizza come personaggio nel racconto, e diventa attore sullo stesso piano di Pigmalione: Galatea non è più solamente il desiderio realizzato di Pigmalione, ma sviluppa i propri desideri e le proprie speranze comportandosi come un essere umano. Ciò che avviene è uno spostamento dell'attenzione dal desiderio tutto umano e univoco dell'uomo di far vivere la propria creazione alle dinamiche di interazione che si instaurano tra i due. Nel racconto del mito di Galatea scritto da Madeline Miller (2013), l'intreccio narrativo è concentrato sul far vedere la doppia natura di statua e donna di Galatea in interazione con Pigmalione, sottolineando i movimenti e gli sguardi, i comportamenti e i gesti, cercando di analizzare le modalità di esercizio del potere interazionale di entrambi i protagonisti. L'evoluzione del mito riesce a mostrarci come, attraverso uno spostamento dell'interesse dal singolo agente alle dinamiche che li coinvolgono entrambi, il modo di considerare l'*altro* in interazione non sia completo se visto solamente dal punto di vista di uno dei due agenti coinvolti; per apprezzarne le caratteristiche e i possibili sviluppi è necessario considerare l'interazione

nel suo complesso, come un'entità autonoma, terza rispetto agli agenti coinvolti, ma che si sviluppa a partire dal contributo dei singoli.

Ora, i robot che abbiamo visto costruiti nel viaggio di andata che abbiamo affrontato, come SHAKEY, sono il corrispettivo novecentesco della statua senza nome: costruiti attraverso l'impiego della tecnologia del tempo, il movimento che si realizza è quello che va direttamente dal costruttore all'artefatto, il quale incorpora lo sguardo del proprio creatore attraverso le teorie sviluppate sul funzionamento della cognizione umana. I problemi riscontrati nella fase di valutazione, però, hanno messo in luce la necessità di tenere presente non soltanto le ipotesi sul funzionamento della mente e l'intento dei costruttori di realizzare robot capaci di funzionare, ma anche e soprattutto il rapporto che la macchina ha con il proprio ambiente e il modo in cui il robot viene valutato dall'osservatore in quanto intelligente e capace di raggiungere i propri obiettivi. Così, la svolta data dalla *Nouvelle AI* e dagli insetti di Brooks illumina gli aspetti relazionali delle prestazioni robotiche, consentendo a questi robot di acquistare la propria autonomia agentiva all'interno dell'interazione, sfruttandola e considerando gli effetti che il comportamento del robot ha sull'essere umano che si trova ad interagire con lui e sull'interazione nel suo complesso. Brooks e i suoi colleghi hanno compiuto il primo passo per costruire robot con una propria autonomia interazionale: Galatee con un nome proprio. Con questa svolta, il robot è arrivato a essere non più solamente l'oggetto delle ambizioni dell'essere umano, creatura nata dal desiderio solitario di costruire forme di vita alternative, ma ha acquistato una propria autonomia diventando un interlocutore nel rapporto con gli esseri umani. Il viaggio di ritorno, dunque, non è che agli inizi: è solo osservando l'interazione con questi agenti, ora capaci di fungere da compagni dialoganti – non solo di suscitare in noi reazioni sociali, ma anche di sfruttarle e concatenarsi a esse – che saremo capaci di capire cosa distingue ancora noi da loro e cosa è possibile migliorare per ottimizzare le modalità di incontro.

Allora, per analizzare il modo in cui una *Galatea moderna*, cioè robot in grado di interagire in qualche modo con noi esseri umani, si presenta agli occhi dell'osservatore, ai nostri occhi, il percorso deve tenere conto: i) del modo in cui i robot imitano gli esseri umani e delle reazioni umane alle loro fattezze estetiche; ii) del ruolo del comportamento del robot nell'interazione; iii) del modo in cui l'interazione si sviluppa concatenando e intrecciando i contributi di tutti gli agenti coinvolti nell'interazione.

In sostanza, vedere il viaggio che abbiamo intrapreso nel capitolo precedente attraverso l'evoluzione del mito di Galatea e Pigmalione ci dà la possibilità di notare almeno due punti fondamentali. In primo luogo, ci consente di riconoscere al robot il ruolo di agente, che attraverso le capacità implementate contribuisce alla strutturazione dell'interazione. Come vedremo in questo capitolo, non sempre il comportamento del robot è considerato come fattore strutturante

dell'interazione. Spesso, l'analisi del contributo del robot si esaurisce guardando alle caratteristiche estetiche che la macchina esibisce, prese in considerazione solamente in relazione alle reazioni interne che suscitano nell'essere umano. L'analisi del comportamento del robot si concentra sovente sugli effetti che questo ha sul sistema cognitivo dell'essere umano, mettendo in secondo piano il contributo all'interazione, considerata per lo più come sistema che accoglie le trasformazioni individuali e che risulta da queste determinata. L'evoluzione del mito, invece, illumina le possibilità che emergono dal considerare insieme il contributo degli agenti alla formazione dell'interazione e il modo in cui questa si sviluppa in modo da influenzare a sua volta il ruolo e le possibilità di azione di ogni agente coinvolto.

In secondo luogo, il mito ci illumina sulla necessità di concentrare la nostra attenzione sul modo in cui l'interazione viene affrontata e gestita da entrambi gli interattanti. Una ripresa moderna del mito di Pigmalione rende ancora più chiaro lo spostamento che intendiamo proporre. Nell'opera teatrale *Pigmalione*, George Bernard Shaw (1913) mette in scena le vicende che coinvolgono il professore di fonetica Henry Higgins, il suo amico il colonnello Pickering e una fioraia, Eliza Dolittle<sup>75</sup>. Il professore scommette che riuscirà a rendere la donna, proveniente da una classe sociale povera e con una scarsa – se non assente – educazione, una signora che sappia muoversi nei salotti dell'alta borghesia inglese, e questo solamente insegnandole a comportarsi seguendo l'accento, l'etichetta e le buone maniere che si confanno a una signora nata ricca. Ciò che viene sottolineato a più riprese nell'opera è che il segreto per far credere alle persone che qualcuno sia capace di fare cose che in realtà non sa fare è di farla comportare esattamente nel modo in cui dovrebbe se avesse quelle stesse competenze. La scommessa che fa il professore con il suo amico consiste nella fiducia data al comportamento. Non importa se Eliza possiede davvero quelle caratteristiche, se le ha introiettate per nascita o se le sono state impartite nell'età adulta e non le sente sue: se sarà capace di dimostrare il comportamento giusto al momento giusto, allora la differenza tra nobili per nascita e poveri alfabetizzati scompare.

Eliza, ennesima Galatea, riesce a essere convincente in un panorama sociale che non le appartiene perché esibisce un comportamento che è simile, del tutto confondibile, con quello delle altre persone: un gioco dell'imitazione, per come lo avrebbe definito Turing (1950; cfr. §5), ottimamente realizzato. L'opera ci instrada, allora, sul contributo che il comportamento esibito dall'agente ha nel complesso dell'interazione. L'osservatore valuta le prestazioni dell'agente a partire dal modo in cui questo si comporta nell'ambiente in cui si muove, con gli agenti con cui si trova a relazionarsi, dimostrando

---

<sup>75</sup> È proprio da Eliza Dolittle, protagonista dell'opera di George Bernard Shaw (1913), che Joseph Weizenbaum prende ispirazione per nominare il suo programma ELIZA (Weizenbaum 1966; cfr. §2.1.1.3, p. 59; §2.3.1.1), dato che riusciva a dimostrare capacità discorsive che potevano far credere all'interlocutore che ci fosse una persona reale dall'altra parte dello schermo, così come Eliza era capace di far credere di appartenere a una classe sociale diversa. Per lo stesso motivo, Douglas Hofstadter identificherà questo fenomeno come effetto ELIZA (Hofstadter 1996).

capacità che ha appreso nel corso delle sue esperienze e raggiungendo obiettivi valutabili per il loro successo o fallimento. Il parallelo con la robotica sociale è diretto: quello che viene sottolineato è il fatto che si possa dimostrare una capacità interazionale efficace, che instaura un regime di comportamento sociale comparabile con gli altri esempi della stessa natura. Coordinarsi con i nostri movimenti, allestire un campo di co-creazione di significato, rendere l'interazione simile ad altre con cui siamo familiari non solo sul piano comportamentale, ma anche su quello estetico: così, Galatea ed Eliza ci ricordano dove fissare i nostri obiettivi, senza però poter fare affidamento sul contributo divino o sulle capacità umane di apprendimento. Nel caso della robotica, si tenta di colmare il divario tra intenti e realizzazioni costruendo un percorso che tenga insieme le idee e le teorie sul funzionamento della mente umana e gli strumenti tecnologici con cui poterle realizzare.

Contributo del robot in interazione, mimesi del vivente e comportamento, allora, saranno gli oggetti di indagine di questo capitolo. Per questo, si cercherà di rendere conto del panorama attuale degli studi della *Human-Robot Interaction* (HRI, §3.2). In seguito, si concentrerà l'attenzione sugli studi inerenti alla cognizione sociale, sia per quanto riguarda i tentativi di implementazione nel robot, sia per spiegare cosa accade all'essere umano di fronte al robot (§3.3). Si procederà delineando una serie di osservazioni sui modelli adottati e, a partire dalla rilevazione di alcuni punti deboli o problematici, si costruirà un paradigma alternativo sulla base delle teorie della cognizione sociale per come elaborate nella corrente enattivista delle scienze cognitive: ciò comporterà l'analisi del modo in cui viene restituita la figura del robot in interazione in quanto artefatto mimetico e sintetico (§3.4). Nel percorso, si arriverà a delineare un quadro in cui la duplice natura dell'artefatto robotico non può che far emergere una *sincreticità*, in cui l'anima sintetica e quella mimetica convivono e non possono che influenzarsi a vicenda (§3.5).

### 3.2 Human-Robot Interaction: questioni e prospettive

Il campo disciplinare degli *Human-Robot Interaction* (HRI) *studies* è un dominio di ricerca ad oggi ben strutturato, il cui obiettivo è esplorare, analizzare, comprendere e migliorare le modalità in cui esseri umani e robot interagiscono (Dautenhahn 2007a; Goodrich e Schultz 2007). Formalmente, la disciplina della HRI è definita come

[...] the science of studying people's behavior and attitude toward robots in relationship to the physical, technological and interactive features of the robots, with the goal to develop robots that facilitate the emergence of human-robot *interactions* that are at the same time efficient (according to the original requirements of their envisaged area of use), but also acceptable to people, and meet the social and emotional needs of their individual users as well as respecting human values. (Dautenhahn 2014, corsivo dell'autrice)

Il primo incontro rivolto a convogliare gli sforzi investigativi di diverse discipline, quali la robotica, le scienze cognitive, l'analisi dell'interazione, la psicologia e la linguistica, avviene nel 1992

con l'inaugurazione del *IEEE International Symposium on Robot & Human Interactive Communication* (RoMan), che da quel momento in poi si sarebbe svolto annualmente (Bartneck et al. 2020). L'esigenza era quella di riunire gli approcci interessati alle tematiche riguardanti l'interazione tra umani e robot, sia sul versante dell'ottimizzazione ingegneristica degli artefatti robotici (Breazeal 2004b), sia nello studio dell'esperienza dell'essere umano di fronte a loro (Kim 2022).

Inizialmente, lo studio della HRI emerge come un campo all'interno degli studi dell'interazione tra esseri umani e computer (*Human-Computer Interaction*, HCI, cfr. Kiesler & Hinds 2004), ma il crescente e repentino sviluppo in ambito tecnologico di robot che potessero implementare capacità relazionali e suscitare in qualche maniera reazioni sociali ha condotto all'autonomizzazione della disciplina intorno all'inizio degli anni '90 (Breazeal 2004b; Dautenhahn 2014; Bartneck et al. 2020). Come abbiamo osservato nei casi dei primi robot con capacità sociali, come Cog e Kismet (§2.3.2.), l'interazione che poteva instaurarsi con queste nuove forme di tecnologia si presentava come molto più collaborativa rispetto a quella tra umani e computer, data la natura situata ed *embodied* dell'incontro tra umani e robot (Dautenhahn 2007a: 103).

Tuttavia, è proprio prendendo in considerazione alcune teorie elaborate nel campo della HCI (a seguito dell'inclusione dei robot, anche detta Human-Machine Interaction, HMI) che l'indagine sulla HRI prende avvio. In particolare, intorno alla metà degli anni '90, la teoria della *Media Equation* si stabilì come ipotesi guida per la comprensione del funzionamento dell'interazione tra esseri umani e macchine tecnologicamente avanzate (Reeves & Nass 1996). L'idea chiave della *media equation* consiste nell'ipotizzare che gli esseri umani si comportino davanti a computer, robot e strumenti tecnologicamente complessi e che dimostrano un sufficiente grado di autonomia esattamente come farebbero di fronte ad altri esseri umani: in sostanza, l'idea è che «media equal real life» (Ibid.: 5), non solo sporadicamente, in circostanze specifiche e in dipendenza alla condizione dell'utente umano, ma pervasivamente, indifferentemente dalle situazioni e dal contesto socio-culturale di riferimento. Le spiegazioni date a supporto di questa teoria si appoggiano a esperimenti che fanno vedere come le risposte sociali, emotive e cognitive dei soggetti sperimentati di fronte a computer siano sovrapponibili a quelle che intercorrono tra umani: è sufficiente che ci siano alcuni *minimal cues* affinché l'agente umano attivi un comportamento sociale e interpreti il comportamento della macchina tecnologica in chiave antropomorfa (Nass et al. 1993; Reeves & Nass 1996; Nass & Moon 2000). È importante notare che questa è stata una delle prime teorie ad essere applicata per l'analisi e la comprensione dell'interazione tra esseri umani e robot, contribuendo a costruire un *framework* interpretativo sulla base di un'uguaglianza tra questo tipo di interazione e quella intrattenuta tra esseri

umani. In sostanza, fin dal principio, l'adozione della teoria della *media equation* ha definito i parametri attraverso cui considerare la HRI sotto la lente dell'interazione interindividuale.

A partire da questo paragone, infatti, il campo di indagine della HRI ha convogliato i suoi sforzi nel voler comprendere come gli studi e le teorie riguardo all'interazione tra umani possano essere applicati a quella tra umani e robot. Riconoscendo, quindi, la necessità di costruire robot che dimostrino una capacità sociale che possa suscitare una reazione simile a quella di fronte a un essere umano, la disciplina della HRI ha prodotto numerosi tentativi di formalizzazione, soprattutto in relazione a due questioni principali.

### 3.2.1 *Che cosa è un robot sociale?*

Ad oggi, la classe dei robot si configura come molto variegata al suo interno. Infatti, chiamiamo robot allo stesso tempo macchine che realizzano ricette in cucina, tagliaerba automatici, bracci industriali automatizzati, alcuni tipi di giocattoli che somigliano ad animali (come gatti, cani e foche) e riproduzioni automatizzate di esseri umani. Il campo di studi della HRI si occupa di come avviene l'interazione con ognuno di questi robot, ma è chiaro che non tutti i tipi di robot che abbiamo elencato necessitano della capacità di interagire attivamente con noi, e ancora minore è il numero di robot che devono mostrare capacità sociali: non c'è bisogno che il mio robot da cucina mi saluti la mattina quando mi faccio il caffè, né che il mio tagliaerba si preoccupi del mio stato d'animo; un braccio robotico adibito alla catena di produzione in una fabbrica non deve (necessariamente) possedere la capacità di capire quando un essere umano è triste. La questione però è diversa se si tratta di robot esplicitamente realizzati per interagire (attivamente o passivamente) con gli umani: robot impiegati nelle terapie (come Paro, Sharkey & Wood 2014, o Kaspar, Dautenhahn et al. 2009), nell'assistenza agli anziani (come Robear, Bedaf et al. 2017), robot costruiti per svolgere lavori in cui è necessario un rapporto con il pubblico (come receptionist, ad esempio ASKA, Nisimura et al. 2002; o camerieri, come Kenon<sup>76</sup>), affinché possano svolgere il compito per cui sono stati costruiti, devono necessariamente possedere e dimostrare alcune capacità sociali. Ma *quali*? Una delle domande su cui la ricerca della HRI ha confluato i suoi sforzi è relativa proprio alle qualità che possono rendere un robot effettivamente sociale. Infatti, i vari gradi di socialità esibiti dal robot non devono essere solamente adatti al contesto in cui il robot viene utilizzato, realizzabili dall'*hardware* del robot su cui vanno implementati e dimostrati in un momento appropriato, ma devono essere integrati in un comportamento unitario che renda l'interazione più fluida possibile (Dautenhahn 2007a, 2007b, 2014).

---

<sup>76</sup>[https://www.keenon.com/en/partnership/index.html?gad\\_source=1&gclid=CjwKCAjwx-CyBhAqEiwAeOcTdSmxgFvQZGuWmXfSYs2dqJA1xzJO1VKayplUAAnq6kB\\_0BIY5N6OCBoC0ysQAvD\\_BwE](https://www.keenon.com/en/partnership/index.html?gad_source=1&gclid=CjwKCAjwx-CyBhAqEiwAeOcTdSmxgFvQZGuWmXfSYs2dqJA1xzJO1VKayplUAAnq6kB_0BIY5N6OCBoC0ysQAvD_BwE)

La concentrazione dell'attenzione su questi aspetti relativi alla socialità esibita dal robot ha condotto a formulare domande circa le caratteristiche essenziali di un agente sociale: quali sono le caratteristiche minime che definiscono un agente sociale come tale? Come possiamo riprodurle, se individuate, nel comportamento del robot?

Con l'obiettivo di iniziare a delineare una serie di risposte alle domande relative alla natura e alla realizzazione delle capacità sociali, alcuni studiosi hanno cercato di sistematizzare e classificare le topologie di robot già esistenti (Breazeal 2002, 2003a; Fong et al. 2003, Dautenhahn 2007b). Il risultato è la formulazione di una gerarchia delle tipologie di robot sociali in ordine crescente di socialità implementata<sup>77</sup>.

- i) Al livello minimo di capacità sociali troviamo i *socially evocative* robot, ovvero robot che fanno affidamento sulla tendenza dell'essere umano ad antropomorfizzare ciò che mostra anche solamente una minima somiglianza all'essere umano e che sfruttano il legame affettivo suscitato da interazioni il cui interlocutore necessita di attenzione e cura. È il caso, ad esempio, dei robot giocattoli come Furby, o del robot Paro, realizzati appositamente per innescare la tendenza umana ad attribuire capacità sociali ai robot e a 'prendersi cura' di loro, cercando di emulare l'interazione tra bambini (o cuccioli) e *care-giver* (Breazeal 2002, 2003a, 2004a). In questo caso, la socialità dell'interazione è sbilanciata dalla parte dell'umano, in quanto il robot non esibisce realmente comportamenti sociali indirizzati a uno scopo ma emette segnali o mostra caratteristiche estetiche con l'obiettivo di stimolare l'essere umano a fare la maggior parte del lavoro.
- ii) Al livello immediatamente superiore, troviamo i robot *socially situated*, che sono capaci di reagire in modo sociale agli stimoli, di riconoscere altri agenti (sia robot che umani) e di interagire con gli oggetti all'interno dell'ambiente in cui si trovano ad agire (Fong et al. 2003). In questo caso, i robot sono capaci solamente di comportamenti passivi: non possiedono dunque complesse abilità sociali, ma comunque dimostrano una sensibilità agli stimoli esterni, compresi quelli che avvengono nell'interazione, dando l'impressione di avere alcune capacità che non sempre possiedono. L'esempio di Kismet e del racconto riportato da Breazeal nel capitolo precedente rende bene l'idea: anche una risposta semplice, interposta tra due frasi dell'interlocutore umano, sebbene non verbale, lo induce a un'interpretazione antropomorfizzante delle azioni del robot, contribuendo a valutare positivamente il suo comportamento e a favorirne l'interazione (Breazeal 2002, 2003a).

---

<sup>77</sup> In Breazeal (2002, 2003a), la terminologia è leggermente diversa: viene usata l'etichetta *social interface* come sottoclasse intermedia tra *socially evocative* e *socially situated* (in Breazeal 2003a definiti *socially receptive*), ma la sostanza non cambia. Si è scelto di propendere per la terminologia data da Fong et al. (2003) e da Dautenhahn (2007b) perché ad oggi più diffusa nel settore.



- iii) I robot *sociable*, oltre a reagire agli stimoli sociali, sono anche capaci di agire proattivamente all'interno del proprio ambiente, stabilendo autonomamente i propri obiettivi relativi alle interazioni sociali e alle emozioni e cercando di soddisfarli attraverso azioni intraprese spontaneamente. In particolare, questo tipo di robot è il primo nella scala che richiede l'esplicita formulazione di una teoria riguardo al funzionamento della socialità (Breazeal 2002, 2003a). Nell'interazione con questa tipologia di robot, l'azione è più collaborativa ed è più bilanciata rispetto a quella con i robot *socially evocative*, dal momento che il robot dimostra qualche grado intraprendenza.
- iv) La tipologia dei robot *socially intelligent* esibisce caratteristiche sociali ancora maggiori, dimostrando capacità di coordinarsi socialmente in uno stile simile a quello dell'essere umano. Ciò è reso possibile da un modello esplicito e dettagliato di cognizione umana e di competenza sociale (Dautenhahn 1998). Un esempio in questo senso è iCub, un robot sviluppato per dimostrare capacità sociali molto simili a quelli dell'essere umano, capace di interagire in modo interattivo (Metta et al. 2010; Parmiggiani et al. 2012).
- v) Infine, i robot *socially interactive* costituiscono il grado più alto della scala in termini di socialità. Per questo tipo di robot, l'interazione è importante non solo nel contesto di riferimento, ma anche per sviluppare una propria 'personalità'. L'interazione è quindi resa indispensabile in un processo di apprendimento che si dispiega in ogni interazione, in cui le azioni vengono testate in un contesto congiunto e in accordo agli obiettivi e alle tendenze dell'agente (Dautenhahn 2007b; Fong et al. 2003).

Oltre ad essere una scala che classifica le tipologie di robot esistenti, quella appena esposta si delinea anche come un percorso programmatico di sviluppo. Infatti, nonostante i tentativi di sviluppare robot con altissimi gradi di socialità, fino ad ora i risultati si sono fermati ai *robot socially intelligent*. Robot con capacità come quelle esposte per la categoria dei *socially interactive* fanno sicuramente parte degli obiettivi a lungo termine della robotica sociale, ma ancora difficilmente realizzabili dal punto di vista tecnologico.

Per alcuni, ciò che si cerca di realizzare, che contraddistingue la tipologia dei robot *socially interactive*, è la capacità di essere un *sostituto* dell'essere umano (Dumouchel e Damiano 2017: 31-63). Il senso, ovviamente, non va inteso come quello messo in scena in molti dei racconti fantascientifici e nelle narrazioni apocalittiche che circolano oggi nella nostra cultura: il sostituto di cui si parla non è un robot che deve andare a rimpiazzare l'essere umano definitivamente, sottraendo possibilità e libertà, ma un'agente capace di svolgere i compiti per cui viene costruito nello stesso modo (o con la stessa efficienza) con cui lo farebbe un essere umano. In questo senso, un *sostituto* è

«qualcuno che sostituisce qualcun altro senza prenderne il posto, cioè senza togliergli la funzione che lo caratterizza» (Ibid.: 38). La domanda, allora, diventa: quali sono le caratteristiche sociali che devono essere dimostrate da un robot per essere considerato un sostituto?

A questo scopo, oltre ai contesti pragmatici in cui la costruzione del robot è indirizzata dalle modalità d'uso che ne verranno fatte, i robot sociali vengono costruiti anche per testare le teorie relative alla cognizione sociale per come avviene negli esseri umani. Ad esempio, Abel (Cominelli et al. 2021) è un robot androide costruito appositamente per indagare i meccanismi di cognizione sociale nell'essere umano, mettendo alla prova il modo in cui le persone reagiscono alla sua presenza e interagiscono con lui. In questo caso, l'adeguatezza delle caratteristiche sociali implementate nel robot non si verifica tanto nel contesto di applicazione pragmatico, ma soprattutto nel funzionamento stesso del robot in interazione con l'essere umano.

In ogni caso, a fronte di questa categorizzazione, va notato che i concetti di robot e, a maggior ragione, di robot sociale non possono essere considerati come definitivi, universali e sedimentati in un significato dizionarioale, ma sono soggetti a un'evoluzione che procede di pari passo allo sviluppo tecnologico e si concatenano alle diverse sensibilità che caratterizzano ogni disciplina in un determinato momento del proprio sviluppo. Ad esempio, per quanto riguarda il concetto di robot, è facile immaginare che il significato che attribuiamo oggi a questo termine sia abbastanza diverso da quello che i ricercatori della SRI usavano per denotare SHAKEY, che non includevano alcun meccanismo di cognizione sociale, e molto diverso da quello che intendeva Čapek, per cui i robot erano simbiotici a metà tra il meccanico e l'organico: questo perché, nutriti dalle numerose narrazioni e conoscendo molti esempi di robot ad oggi realizzati, la nostra concezione di robot si è evoluta fino a denotare, ad oggi, una macchina molto complessa, che in qualche modo riesce a interagire con noi, che può aiutarci in qualche compito, che può anche spaventare, ma che nei contesti reali sembra statisticamente risultare utile. Come affermato da Dautenhahn (2014), «The concept of robot is a *moving target*, we constantly reinvent what we consider to be 'robot'». Allo stesso modo, anche l'aggettivo "sociale" accostato alla robotica si è modificato nel corso del tempo: se, come nota Breazeal (2003b: 168), agli albori della disciplina il termine definiva sistemi multi-agente, in cui più robot interagivano tra di loro, negli ultimi venti anni è passato a denotare processi fortemente associati al comportamento sociale tipico degli esseri umani.

Nonostante la temporaneità delle definizioni, un'isotopia riguardo le caratteristiche che un robot sociale deve avere necessariamente per essere considerato tale è stata rintracciata: deve essere capace di interagire con noi nell'ambiente che abitiamo, anche in forma minima. Come individuato da un'analisi sulle varie definizioni utilizzate in campo scientifico e nella cultura popolare e

fantascientifica, la caratteristica della socialità è rintracciata nella capacità di interazione fluida con gli esseri umani (Sarrica et al. 2019).

A partire da qui, quindi, le domande che si configurano riguardano principalmente come identificare, descrivere e spiegare l'interazione che si instaura tra gli esseri umani e questo tipo di agenti robotici.

### 3.2.2 *Come si dà l'interazione tra robot ed esseri umani?*

I tentativi di comprensione provenienti dalle diverse branche disciplinari coinvolte nella HRI cercano tutti di dare una risposta a questa concisa quando complessa domanda. A partire dalla rilevazione di un certo grado di socialità del robot, l'obiettivo è capire come l'interazione si articola attraverso diverse concettualizzazioni del robot come agente sociale: il robot è un agente al pari dell'essere umano? O ci sono nuove forme attraverso cui è possibile inquadrare il suo contributo all'interazione?

Un primo orientamento pone l'enfasi sulle qualità che avvicinano il robot all'essere umano, sulla scia iniziata dalla *media equation*, considerando lo studio della HRI come un regime paritario, in cui cioè è possibile mettere sullo stesso piano il contributo umano e quello robotico. In questo senso, l'indagine si concentra nel produrre evidenze sperimentali che verifichino la similarità (o l'uguaglianza) delle reazioni corporee e cognitive nell'interazione tra umani e umani a quella tra umani e robot. Ad esempio, una serie di esperimenti mostra come alcuni dei meccanismi cognitivi e dei pattern neurali associati a funzioni cognitive di natura sociale nell'interazione tra umani siano attivati anche nella HRI. Tra questi, la *risonanza motoria*, ovvero l'attivazione del sistema di controllo motorio dell'osservatore durante la percezione delle azioni dell'altro in interazione (Sciutti et al. 2012a), e degli schemi di cooperazione o competizione a questa collegati (Bisio et al. 2014), la percezione dell'azione dell'altro come orientata a uno scopo (Sciutti et. al 2012b; Sciutti et al. 2013), la comprensione delle azioni dell'altro in ottica sociale (Sciutti et al. 2015), l'attribuzione di intenzionalità alle azioni del robot in interazione (Wiese et al. 2017; Perez-Osorio & Wykowska 2019, 2020; Marchesi et al. 2019, 2021) sono state considerate come evidenze della possibilità di trattare il robot come agente sociale alla pari dell'essere umano<sup>78</sup>. In questo quadro, quindi, il robot è validato come agente sociale a tutti gli effetti, capace di interagire con l'essere umano ad un livello sufficiente di socialità, attivando gli stessi meccanismi di cognizione sociale.

D'altra parte, considerando invece il robot come un nuovo tipo di agente, vicino ma comunque differente dall'essere umano, ciò che viene enfatizzata è la somiglianza del robot ad altri tipi di agenti

---

<sup>78</sup> Si parlerà approfonditamente di questi esperimenti in §4. Per una panoramica degli studi neuroscientifici in questo senso, si veda Chaminade & Cheng (2009); Cross et al. (2019); Cross & Ramsey (2021).

con cui interagiamo. In questa direzione si muovono i contributi che identificano i robot come partner interazionali simili ad animali con capacità sociali (Prescott 2017; Shaerer et al. 2009; Darling 2021). Sotto questa prospettiva, l'interazione tra robot e umani non potrebbe essere assimilata né a quella con gli umani, dal momento che non dimostrano le stesse posizioni etiche nei loro confronti rispetto a quelle che adotterebbero in contesti di interazione intraspecifici (cfr. Prescott 2017), né come semplici macchine, dal momento che ai robot vengono riconosciuti più responsabilità rispetto a computer e tagliaerba e, di conseguenza, anche una serie di diritti (Gunkel 2012, 2014; Hogan 2017). Come argomentato da Coeckelbergh (2010, 2011), sia gli animali sia i robot si configurano come *altro* in interazione rispetto agli esseri umani, stabilendo una continuità nella loro diversità dall'agente umano. Non troppo umani, non troppo macchine, quindi, i robot instaurerebbero un tipo di interazione che è molto più simile a quella che abbiamo con i nostri animali domestici, per cui è da questo tipo di rapporto che dobbiamo partire per analizzare la HRI (Darling 2021).

Ancora diverso è l'approccio alla HRI che intende il robot solamente come uno speciale e complicatissimo tipo di macchine: macchine che sanno parlare, a volte, che sanno coordinarsi ai nostri movimenti, che danno l'impressione di essere capaci di comprenderci, ma pur sempre macchine. Per questo, il discorso relativo ai valori morali o etici presenti negli altri orientamenti è qui più sporadico, quando non completamente assente, nutrendo anzi considerazioni circa la natura a-morale e non-etica del robot come agente (Bryson 2010). Qui, la HRI è vista ancora quasi esclusivamente come una branca della HCI, e lo studio dell'interazione diventa oggetto di classificazioni e tassonomie al fine di incasellarla in uno dei regimi interazionali che l'essere umano mette in pratica con le macchine. Tra questi tentativi, alcuni si focalizzano sulle reazioni dell'essere umano nell'interazione (Scholtz 2003; Ji et al. 2020; Cross & Ramsey 2021), mentre altri si interessano del sistema interazionale nel suo complesso (Goodrich 2004; Schmidtler et al. 2015), delle funzioni necessarie al robot per poter interagire con l'essere umano (Bartneck & Forlizzi 2004; Beer et al. 2014), o sui criteri di valutazione dell'interazione stessa (Jost et al. 2020; Veling & McGinn 2021). Tuttavia, nonostante il grande lavoro che è stato fatto negli ultimi vent'anni per cercare di inquadrare i vari approcci per lo studio della HRI, spesso le tassonomie non portano poi a uno sviluppo della disciplina e degli approcci che la costituiscono, ma rimangono ancorate a livello teorico come esercizio di stile (cfr. Yanco & Drury 2002, 2004;). Come sottolineato da Onnasch & Roesler (2021), la mancanza di uniformità e generalizzabilità nel settore è uno dei problemi principali dell'organizzazione tassonomica e classificatoria della HRI, il cui obiettivo è stabilire una classificazione definitiva che tenga conto delle caratteristiche del robot, delle azioni e reazioni dell'essere umano e del sistema interazionale in sé. Considerando però le riflessioni riportate riguardo alla temporaneità delle definizioni degli agenti e delle categorie coinvolte, l'obiettivo di arrivare a una classificazione definitiva risulta più difficile di

quanto immaginato. Va inoltre notato che questo tipo di approccio spesso non consente di inquadrare lo studio della HRI attraverso una delle sue componenti costitutive: l'interazione. In questi studi, infatti, il vero e proprio svolgimento dell'interazione non è preso in considerazione. L'attenzione è invece puntata sugli effetti che possono essere estratti dall'incontro dei due tipi di agenti, di conseguenza sottovalutando o tralasciando completamente il ruolo che la componente corporea e in tempo reale ha sui possibili risultati.

In sostanza, per rispondere alla domanda circa le modalità di darsi dell'interazione tra umani e robot, nell'attuale panorama della HRI si ricorre a diverse forme di analogia tra l'interazione in oggetto e altre forme di interazione che l'essere umano realizza con altri tipi di agenti. Il robot è alternativamente un'altra forma di umanità, una forma di animalità o una macchina molto complessa, che a volte può confonderci. Come è evidente, i vari tentativi di comprensione, ancora oggi, risultano in piena evoluzione: nessuna soluzione definitiva è stata raggiunta, e nemmeno una definizione condivisa dei termini chiave della disciplina. Robot, interazione ed essere umano restano ancora termini vaghi, aperti all'interpretazione e alle modalità di indagine che le varie discipline intente nello studio della HRI intraprendono per comprendere meglio il loro oggetto. Dal punto di vista delle discipline e delle teorie, infatti, il campo di studi risulta altamente frastagliato al suo interno, comprendendo approcci provenienti da diversi campi disciplinari che hanno ritagliato la propria indagine dell'interazione tra umani e robot a partire dai presupposti teorici su cui erano basate e rivolgendo l'attenzione verso i possibili punti di congiunzione.

### 3.2.3 *Un dominio eterogeneo*

Guardare al modo in cui le diverse discipline si compongono per costruire il campo disciplinare della HRI e contribuiscono a delineare i confini e gli oggetti di analisi, allora, ci consente di fare chiarezza sul modo in cui la nostra indagine può posizionarsi nel dibattito e di far emergere alcune questioni che ad oggi sembrano passare sottotraccia.

Come sottolineato già, gli studi sulla HRI sono effettuati a partire da diversi domini disciplinari, che portano con sé ognuno la propria metodologia, i propri parametri di giudizio e le proprie definizioni dei termini in gioco. Questo rende molto complesso il processo di sistematizzazione dei contributi che provengono da prospettive anche molto diverse tra loro. Un tentativo di classificazione delle prospettive che compongono la HRI è stato fornito da Dautenhahn (2007b) e ad oggi è quello maggiormente adottato per la classificazione dei contributi nel dominio.

In primo luogo, la prospettiva della *Robot-centred HRI* si concentra nell'indagine del robot «as a creature, i.e. an *autonomous entity* that is pursuing its own goals based on its motivations, drives, and

emotions» (Ibid.: 683). In questo senso, si tratta di concentrare l'attenzione su cosa sia necessario al robot per essere efficace, e ciò significa considerare l'interazione come un mezzo per sopperire a determinati requisiti che il robot richiede. Questo si concretizza nello sviluppo di controlli sensorimotori, modelli di comportamento e sistemi di regolazioni emotivi e motivazionali, i quali regolano a loro volta l'interazione nell'ambiente sociale per estrarne ciò che viene impostato come pertinente e significativo dal costruttore. Per questa prospettiva, l'attenzione è dunque rivolta soprattutto all'*hardware* del robot e alle possibilità che vengono consentite dal suo interfacciarsi nell'ambiente (Kobayashi & Kuno 2010; Kanda & Ishiguro 2017).

In secondo luogo, l'approccio *Human-centred* alla HRI raccoglie quelle prospettive che indagano il modo in cui il robot può risultare «*acceptable and comfortable to humans*» (Dautenhahn 2007b: 684). L'attenzione è quindi rivolta alle reazioni dell'essere umano di fronte al robot, soprattutto in merito all'aspetto o al comportamento mostrato dalla macchina. I principali interessi di questo approccio si concentrano nel i) trovare un equilibrio tra aspetto del robot e capacità comportamentali; ii) comprendere quali siano i livelli di accettabilità del comportamento del robot (a volte elaborando anche linee guida per un'etichetta, o *robotiquette*, Ogden & Dautenhahn 2000); iii) sviluppare nuove metodologie per lo studio e per la valutazione della HRI; iv) individuare esigenze personali o collettive nell'interazione; v) evitare effetti spiacevoli che possono mettere a rischio l'interazione, come l'*Uncanny Valley Effect* (Mori 1970; MacDorman & Ishiguro 2006; cfr. §5.2). Non c'è, in questa prospettiva, esplicito interesse verso le architetture cognitive che permettono al robot di comportarsi in un certo modo, tanto che spesso per gli esperimenti viene utilizzata la modalità di controllo teleoperato, anche detta *Wizard of Oz* (Steinfeld et al. 2009; Riek 2012), senza che il soggetto dell'esperimento sappia che il robot sia controllato da un altro essere umano<sup>79</sup>. Ciò garantisce una maggiore complessità di interazione, dato che il robot funziona in questo caso solamente come interfaccia di comunicazione con l'utente umano, ma permette di sperimentare situazioni altrimenti irrealizzabili. Visto che i meccanismi sociali complessi necessari per un'interazione fluida e in tempo reale sono ancora lontani dall'essere realizzati, in questo modo possono comunque essere simulati creando scenari in cui il robot viene dotato di capacità sociali più sviluppate<sup>80</sup>.

Infine, l'approccio *Robot-cognition centred* considera il robot in quanto sistema intelligente, ovvero una macchina costruita implementando processi di intelligenza artificiale che «*makes decisions on its own and solves problems it faces as part of the tasks it needs to perform in a particular application domain*» (Dautenhahn 2007b: 684). È qui che la robotica sociale sviluppa la sua

---

<sup>79</sup> Sulle interessanti conseguenze dello svelare il controllo teleoperato sul robot da parte di un altro essere umano al soggetto partecipante all'esperimento, si veda Nasir et al. (2022).

<sup>80</sup> Ciò potrebbe essere annoverato tra gli esempi di un'*etica sintetica* per la HRI, in cui vengono testati scenari di interazione che ad oggi non possono essere testati con robot autonomi, ma che potrebbero verificarsi in futuro (Damiano & Dumouchel 2020; Bianchini et al. 2023)

componente cognitiva, prendendo il nome di *cognitive robotics* (Sandini et al. 2021; Cangelosi & Asada 2022) ed elaborando architetture<sup>81</sup> che possano realizzare e sostenere una complessità interazionale compatibile con l'essere umano, oltre che affrontare problemi legati all'apprendimento e al *problem solving*. Questo approccio si è dimostrato essere molto fertile negli anni, con lo sviluppo di molteplici architetture cognitive per l'interazione umano-robot che vanno da modelli più bio-ispirati (cfr. Goertzel et al. 2010; Vernon 2022) o relativi a caratteristiche specifiche del sistema cognitivo umano o del suo comportamento (Baxter et al. 2013), non di rado elaborate in connubio (come per l'architettura di iCub, Vernon et al. 2007)<sup>82</sup>.

Come sottolineato da Dautenhahn, le tre prospettive qui delineate non sono mutualmente esclusive e anzi dovrebbero auspicabilmente completarsi l'una con l'altra al fine di raggiungere una copertura ideale di tutto il ciclo di sviluppo del robot (Dautenhahn 2007b: 684). Tuttavia, nella maggior parte dei casi, le ricerche in ognuna di queste prospettive sono perseguite autonomamente, ed è difficile che si instauri un dialogo tra approcci che costruisca un circolo virtuoso di reciproci arricchimenti. Ciò accade perché i diversi approcci contenuti in ognuna delle prospettive non di rado vedono la HRI solamente come un ulteriore oggetto di analisi da testare dall'interno della propria disciplina, con i propri metodi e criteri, nella maggior parte dei casi senza creare risultati generalizzabili e condivisibili in un approccio congiunto al problema dell'interazione tra umani e robot. Quello che accade è quindi, spesso, che ogni disciplina afferente a una delle tre prospettive porti avanti la propria indagine, confrontando i risultati principalmente al proprio interno e contribuendo alla comprensione della HRI senza cercare di inserire il proprio sforzo all'interno di un quadro più ampio. Questo aspetto è ciò che rende, ancora oggi, gli studi di HRI non una disciplina dai confini e dalle metodologie definiti, ma un campo di studi multi- e interdisciplinari, in cui ognuno degli approcci e delle discipline che si interessano agli studi sui rapporti tra umani e robot apporta il proprio contributo alla comprensione locale del fenomeno osservato. In questo quadro, soprattutto le discipline di matrice umanistica hanno contribuito sostanziosamente all'avanzamento della comprensione di alcuni aspetti della HRI, ma senza sentire come necessaria l'adesione allo stato dell'arte della robotica.

Ad esempio, la sociologia si è inserita nella HRI con un'indagine circa il ruolo del robot come nuovo agente nel panorama socioculturale contemporaneo in linea con gli studi ecologici sull'impatto delle nuove tecnologie (van Oost & Reed 2011; Meister 2014). In questi studi, l'obiettivo è

---

<sup>81</sup> Un'architettura cognitiva robotica è «a software framework that integrates all the elements required for a system to exhibit the attributes considered to be characteristic of a cognitive agent» (Vernon 2022: 193), e costituisce «a concrete framework for more detailed modeling of cognitive phenomena, through specifying essential structures, divisions of modules, relations between modules, and a variety of other aspects» (Sun 2004: 341). È importante nella modellazione cognitiva a livello computazionale perché rende esplicito l'insieme di assunti e presupposti su cui si fonda il modello cognitivo (Vernon 2014; Vernon et al. 2016).

<sup>82</sup> Per una panoramica sulle architetture cognitive e sulle varie applicazioni, compresa la HRI, si veda Kotseruba & Tsotsos (2020).

comprendere come i robot sociali si inseriscano come ulteriore variabile nel complesso meccanismo di costruzione della sfera sociale umana, facendo riferimento a teorie che vedono il contributo degli oggetti e delle nuove tecnologie al pari dell'umano nella costituzione della imponente rete sociale in cui siamo immersi (come ad esempio l'*Actor Network Theory*, ANT, cfr. Latour 2005). Studi come quelli condotti da Sherry Turkle hanno contribuito enormemente a consolidare il contributo della sociologia alla HRI, soprattutto impostando uno sguardo su come considerare i robot in quanto agenti sociali. In particolare, a partire dalla spinta a considerare le nuove tecnologie non solo come oggetti che adoperiamo ogni giorno, ma soprattutto come parte costitutiva della nostra vita sia privata che collettiva, Turkle instrada l'indagine sull'interazione sugli effetti che questa ha sulla persona (Turkle 1984), evolvendosi nel tempo (1995) e modificando anche il modo in cui interagiamo con gli altri (2007; 2011).

Anche la filosofia, nel corso degli anni, ha cominciato ad interessarsi allo studio della HRI, ponendo domande circa la natura dell'interazione con i robot e la possibilità di comprenderli come nuovi partner in intersoggettività. In particolare, la filosofia viene messa in pratica i) sviluppando riflessioni etiche circa la capacità del robot di essere all'altezza delle aspettative in contesti sociali (Nyholm 2020; 2023; Floridi et al. 2018); ii) analizzando le possibilità di interpretare l'agentività, ovvero la capacità del robot di agire autonomamente nel mondo, su diversi livelli (Hakli & Seibt 2017); iii) cercando di comprendere quali siano le caratteristiche che fanno la differenza tra l'essere umano e il robot e interpretando i risultati sperimentali prodotti da altri approcci (Fuchs 2021); iv) proponendo un approccio fenomenologico all'esperienza dell'essere umano di fronte al robot (Coeckelbergh 2010; Bisconti & Carnevale 2021; Fuchs 2021); v) osservando le somiglianze e le differenze che intercorrono a livello culturale nell'interazione, stabilendo un legame con gli studi antropologici e etnografici (Veling 2017; Arora et al. 2024). In altre parole, i contributi filosofici alla HRI negli ultimi anni si sono moltiplicati, contribuendo enormemente a far avanzare lo stato dell'arte sui concetti di interazione con forme alternative di agentività, di agente sociale, delle differenze e delle similarità che ad oggi il confronto fra esseri umani e macchine ci portano a vedere, soprattutto se messi a interagire.

Tuttavia, dato forse il disinteresse diffuso per la componente applicata della robotica, spesso gli studi in ambito umanistico fanno riferimento a versioni idealizzate dei robot, concentrandosi di più sull'impatto che un robot senziente, cosciente e capace di interagire nel mondo ad un livello di complessità uguale o addirittura superiore (ad esempio, con capacità di lettura della mente) a quello dell'essere umano possa avere sul tessuto sociale, politico, etico e culturale e sulle categorie che utilizziamo per comprendere il mondo e l'esperienza. Un esempio emblematico in questo senso sono gli esperimenti mentali di Nick Bostrom (2014), che fanno leva più sull'immaginario fantascientifico



che nutre la nostra considerazione degli agenti artificiale che a una vera e propria considerazione di cosa i robot siano oggi, allo scopo di indagare il modo in cui prospettive filosofiche considerate irrealizzabili possano essere affrontate in un futuro, prossimo, remoto o irrealizzabile. In sostanza, nonostante questi approcci (sociologico e filosofico) contribuiscano al costruire forme di sapere circa il mondo dei robot e dell'interazione con essi, la necessità che si rende palese è quella di fare i conti con le condizioni tecniche e ingegneristiche di costruzione di un robot e con lo stato attuale del progresso tecnologico in questo settore se lo scopo è quello di nutrire delle riflessioni in merito alle contemporanee e prossime possibilità di interazione.

In questo senso, la robotica rimane ovviamente la disciplina maestra che ha trovato nell'HRI un terreno fertile per sviluppare robot con capacità sociale. Infatti, con l'evoluzione degli obiettivi in robotica, cioè con il passaggio dalla modellazione della navigazione e manipolazione dell'ambiente alla necessità di costruire robot capaci di interagire con gli esseri umani, un settore che sistematicamente si occupa di comprendere come sia possibile ottimizzare le indicazioni per la costruzione del robot, guardando alle aspettative dell'agente umano, alle sue paure e alla naturalezza del suo agire sociale, risulta fondamentale per non andare a tentoni e dissolvere in un solo prototipo centinaia di migliaia di dollari di investimenti (Kanda & Ishiguro 2017). Si può dire che il rapporto tra robotica e HRI corre su un doppio binario: da una parte, la robotica prende in consegna gli studi riguardanti la HRI in merito a capacità di interazione, aspettative, dimensione emotiva e cognitiva per sfruttarli nella costruzione di robot e delle loro architetture informate dalle evidenze prodotte nel settore di pertinenza; dall'altra, la HRI mette alla prova le costruzioni che dalla robotica provengono, per vedere come l'interazione si dà nel contesto e quali siano le particolarità che il robot mette in scena. Si costruisce così un rapporto di mutua collaborazione, fondamentale per innescare un circolo virtuoso di scambi teorici e applicazioni in cui una disciplina nutre l'altra con le proprie ricerche (Breazeal et al. 2016). Negli anni, la collaborazione tra robotica sociale e HRI è stata più che fruttuosa (Breazeal 2003, 2004a; Dautenhahn 1999, 2014), e ad oggi si rivela come indissolubile: senza la HRI, robot sociali come ad esempio NAO<sup>83</sup> (Shamsuddin et al. 2011) non avrebbero potuto essere realizzati e di conseguenza impiegati in numerosi e diversi task, dall'assistenza agli anziani (Recio et al. 2013; Pino et al. 2020) e ai bambini con sindrome dello spettro autistico (Shamsuddin et al. 2012; Tapus et al. 2012) allo studio dell'interazione tattile per veicolare emozioni (Andreasson et al. 2018) e al loro riconoscimento (Erden 2013; Valagkouti et al. 2022)<sup>84</sup>.

---

<sup>83</sup> NAO è attualmente una delle piattaforme robotiche più utilizzata al mondo (<https://unitedrobotics.group/en/robots/nao>). La sua piattaforma è completamente programmabile, caratteristica che consente una estesa flessibilità di applicazione in compiti anche molto diversi tra loro.

<sup>84</sup> Per una panoramica completa sulle varie applicazioni di NAO e per il diffusissimo uso in contesti pubblici e di ricerca si veda Amirova et al. (2021).

In sostanza, il legame tra robotica e HRI è uno dei pochi che possa dirsi reciproco: nel caso della sociologia o della filosofia, molto spesso le riflessioni sviluppate rimangono confinate all'interno della propria area di ricerca, senza sortire effetti considerevoli sulle modalità di progettazione e costruzioni di robot pensati per interagire con gli esseri umani. Tuttavia, se alle discipline umanistiche manca spesso un appiglio realistico alle questioni concernenti le capacità attuali dei robot, alla robotica a volte manca una capacità interpretativa che riesca a inserire i propri risultati all'interno di un panorama più ampio e complesso, fatto di dinamiche sociali e culturali.

Si potrebbe dire che il rapporto della HRI con le discipline che la compongono è un rapporto che parte da – e non arriva a – l'interazione tra umani e robot: l'interazione è, nella maggior parte dei casi, considerata come un oggetto da indagare a partire da una disciplina, che testa la propria metodologia di ricerca sul rapporto tra umani e robot senza però istaurare un dialogo con le altre anime che compongono questo campo di studi così interdisciplinare, e contribuire con ciò a generalizzare i risultati all'interno di un quadro più ampio. Ciò è dovuto in larga parte alla divergenza di interessi: in sociologia, l'obiettivo ultimo è costruire un quadro in cui sia possibile comprendere il ruolo del robot come agente nel contesto socio-culturale; per la filosofia, l'obiettivo è sviluppare riflessioni sui concetti che sono implicati in e gravitano attorno all'interazione tra umani e robot, concettualizzando l'esperienza per renderla intelligibile e per sviluppare previsioni sul futuro dei nostri rapporti con gli agenti artificiali; per la robotica, lo studio della HRI è funzionale alla ottimizzazione del design del robot, consentendole di aggiustare i parametri che informano la costruzione con test e esperimenti condotti in interazione con l'umano.

#### *3.2.4 Minimo comun denominatore: l'indagine della socialità*

C'è da constatare, tuttavia, che, a prescindere dallo sguardo disciplinare a cui la HRI viene sottoposta, un ruolo centrale è ricoperto dagli studi che riguardano la socialità dell'essere umano. Indistintamente dalla disciplina di partenza, infatti, il focus sulle modalità in cui l'essere umano si relaziona in modo sociale al robot è di interesse a tutte le indagini. Fin dal periodo di definizione della disciplina, l'attenzione alla natura sociale dell'essere umano ha costituito un punto cardine dell'indagine sulla HRI. Già Dautenhahn (1994, 1995, 1998, 1999, 2004, 2007b, 2014) sottolinea l'importanza di guardare agli studi relativi alla socialità dell'essere umano sia in contesti di interazione intraspecifica, sia quando l'interlocutore è un animale o una macchina. Sottolineando la complessità della sfida di raggiungere il livello di intelligenza manifestato dagli esseri umani<sup>85</sup>,

---

<sup>85</sup> È interessante notare che Dautenhahn (cfr. 2007b, 2014) rivolge la critica riguardo il necessario spostamento di prospettiva, dall'intelligenza generale a quella sociale, sia alla ricerca sull'intelligenza artificiale di stampo cognitivista, sia alle ricerche di Brooks della Nouvelle AI. In questo senso, la critica di Dautenhahn è rivolta non tanto ai presupposti teorici su cui la Nouvelle AI si poggia, quanto all'obiettivo di replicare l'intelligenza umana nella sua generalità.

Dautenhahn sposta il focus distintivo dell'indagine dell'AI, concentrata sulla replicazione di meccanismi di intelligenza generale (cfr. §2.1), alla necessità di comprendere prima di tutto l'intelligenza sociale come chiave per raggiungere capacità più vicine a quelle umane nei robot:

An alternative viewpoint towards AI [...] is to propose that one particular aspect of human intelligence, namely *social intelligence*, might bring us closer to the goal of making robots smarter (in the sense of more human-like and believable in behavior); the social environment cannot be subsumed under 'general environmental factors', i.e. humans interact differently with each other than with a chair or a stone. (Dautenhahn 2007b: 681).

Per portare avanti questa tesi, Kerstin Dautenhahn si rifà a una teoria elaborata dall'antropologo e psicologo dello sviluppo Robin Dunbar (1998, 2023), basata sulla *social intelligence hypothesis*: in sostanza, si ipotizza che l'intelligenza dei primati superiori si sia evoluta principalmente in adattamento alla complessità sociale dell'ambiente, ovvero per interpretare, predire e manipolare i conspecifici. L'intelligenza oggi mostrata dall'essere umano sarebbe dunque il risultato dell'evoluzione di meccanismi cognitivi prima di tutto sociali, sviluppati innanzitutto per affrontare la complessità dell'ambiente popolato da altri individui della stessa specie, solamente in seguito trasferiti o riadattati a meccanismi cognitivi non-sociali come il *problem-solving*.

Nel campo della progettazione di robot che dimostrino un'intelligenza al pari di quella umana, ciò significa non considerare la componente sociale solamente come un'aggiunta o una caratteristica accessoria da implementare, ma come una componente fondativa e basilare del processo di replicazione dell'intelligenza umana:

Applying the 'social intelligence hypothesis' to AI implies that social intelligence is a key ingredient of human intelligence and, as such, a candidate prerequisite for any artificially intelligent robot. Research on intelligent robots usually focuses first on making robots cognitive by equipping them with planning, reasoning, navigation, manipulation and other related skills necessary to interact with and operate in the non-social environment, and then later adding 'social skills' and other aspects of social cognition. [...] to phrase it differently, developing an intelligent robot means developing first a socially intelligent robot. (Dautenhahn 2007b: 682)

Tra tutte le incertezze e le ambiguità, quindi, una caratteristica si staglia come imprescindibile nel quadro di indagine intrapreso dalla HRI: la socialità. Sia dal punto di vista del robot, che deve esibire le giuste e adeguate capacità di interazione sociale in interazione, sia dal punto di vista dell'essere umano, la cui natura sociale viene messa alla prova di fronte a questo particolare tipo di agenti sociali, la socialità diventa un elemento cruciale da indagare per avanzare nella ricerca di tutti i campi della HRI che si occupano di indagare la comunicazione diretta tra umani e robot. Sviluppare un robot che sia prima di tutto un agente capace di cognizione sociale, dunque, diventa il proposito programmatico della ricerca in robotica sociale, in cui la HRI svolge un ruolo fondamentale data proprio la sua fortissima componente interdisciplinare, capace di fornire soluzioni già sviluppate all'interno di altri domini.

Più di tutti, il dominio a cui si è fatto riferimento per spiegare cosa renda l'animale umano un animale sociale è quello delle scienze cognitive. Come abbiamo visto, le scienze cognitive sono a loro volta composte da diverse anime, che comprendono la psicologia, la filosofia, le neuroscienze da un lato e la robotica, l'AI e le scienze dell'artificiale dall'altro. Fin dagli albori negli anni '60, le scienze cognitive hanno indagato il modo in cui l'individuo interagisce e comunica con i suoi simili, cercando di trovare spiegazioni a quella che sembra essere una caratteristica fondamentale dell'essere umano<sup>86</sup> (Hilton 2012; Boden 2006: 298-316). Ma, tra le diverse ipotesi elaborate in questo dominio per dare conto della componente sociale della cognizione umana, negli studi di HRI una precisa teoria riguardo il funzionamento della cognizione sociale è entrata a far parte del bagaglio interpretativo di tutte le discipline coinvolte nell'indagine dell'interazione tra umani e robot: la *Theory of Mind*. Questo è accaduto non perché si sia applicata perizia nella scelta del modello di cognizione sociale più adeguato a spiegare ciò che accade nell'interazione tra umano e robot, ma perché la *Theory of Mind* si è attestata fin dagli anni '80 come la principale e più adottata modalità di spiegazione della cognizione sociale in qualsiasi sua forma.

### 3.3 Scienze cognitive, cognizione sociale e HRI

La *Theory of Mind* (ToM), o Teoria della Mente, è una teoria elaborata in psicologia che identifica la capacità di ragionare sui propri stati mentali (ovvero credenze, desideri, intenzioni, conoscenze, ecc.) e, allo stesso tempo<sup>87</sup>, di attribuire stati mentali analoghi agli altri, allo scopo di predire e spiegare il loro comportamento (Apperly 2011, 2012; Baron-Cohen 1995). Sviluppata a partire dagli anni '80, questa teoria ha acquisito sempre più popolarità, fino a diventare il principale riferimento per spiegare il modo in cui gli esseri umani (e primati superiori) comprendono gli altri (Leslie et al. 2004; Baron-Cohen 1999). Per capire meglio il peso teorico della ToM negli studi della cognizione sociale e la sua rilevanza nella HRI, dobbiamo innanzitutto comprendere quali siano le basi su cui si fonda e quali siano i meccanismi considerati fondamentali per il suo corretto funzionamento.

La teoria della mente si sviluppa a partire da alcuni studi etologici sul comportamento dei primati superiori in contesti sociali (Premack & Woodruff 1978) e viene importata nel dominio delle scienze cognitive da filosofi (Bennett 1978; Dennett 1978b) e psicologi (Pylyshin 1978). Entrando in contatto

---

<sup>86</sup> Per una panoramica sulla nascita della psicologia sociale e del suo ruolo nelle scienze cognitive, cfr. Kruglanski & Stroebe (Eds., 2011).

<sup>87</sup> Tali abilità proprie della ToM sono classificate come di primo e secondo ordine. La ToM di primo ordine identifica l'abilità di inferire gli stati mentali degli altri (ad esempio, nella forma "io credo che tu creda"), mentre la ToM di secondo ordine concerne la capacità di inferire stati mentali di un'altra persona in relazione a una prospettiva terza (ad esempio, nella forma "io credo che tu creda che egli crede") (Duval et al. 2011). In letteratura, si considera la ToM di primo ordine come più fondamentale e acquisita nel corso dello sviluppo primario, mentre la ToM di secondo ordine si acquisisce in fasi di sviluppo successive e sulla base della prima, venendo considerata come più complessa e in termini meta-rappresentazionali (Morin 2006).

con le teorie computazionaliste e funzionaliste di maggior rilievo, la teoria delle mente, già improntata su basi logiche e inferenziali, si salda ancora di più alla metafora della mente come macchina e della cognizione umana come processore di informazioni, arrivando a delinearsi come teoria computazionale della cognizione sociale. Ciò è testimoniato anche nelle più recenti elaborazioni della ToM, in cui il principio cardine della teoria della mente è descritto funzionare esattamente come un'euristica, composta da un processo di selezione e un meccanismo che ne consente l'applicazione (Leslie et al. 2004: 531). In sostanza, seguendo l'approccio computazionalista alla cognizione, anche il processo di comprensione delle azioni degli altri è qualcosa da ricondurre alla natura proposizionale, linguistica e logica dei meccanismi cognitivi (Apperly 2011).

Ancora oggi, infatti, il meccanismo cognitivo principale su cui la ToM si basa è quello dell'inferenza: per comprendere gli stati mentali dell'altro, l'individuo deve operare inferenze sulla base: i) del contesto condiviso, ii) degli indizi percettivi e iii) dell'interpretazione delle azioni dell'altro, causate dai suoi stati mentali (Byom & Mutlu 2013). Per decretare la necessità di meccanismi di ToM per comprendere le azioni dell'altro, Baron-Cohen (1995) sviluppa un esperimento mentale molto semplice: si immagina un individuo che entra improvvisamente in una stanza, si guarda un po' intorno e poi esce. Come si può interpretare il suo comportamento se non facendo riferimento a una serie di inferenze circa i suoi stati mentali? Probabilmente, suggerisce Baron-Cohen, elaborare spiegazioni circa il suo comportamento è proprio quello che faremmo: potremmo pensare che abbia dimenticato qualcosa nell'altra stanza, oppure che stia cercando qualcosa e si sia ricordato che quello che cercava era da qualche altra parte. Elaboreremmo, cioè, una spiegazione basata sull'osservazione del suo comportamento che ci possa rendere conto dei motivi che lo hanno spinto a compiere queste azioni, inferendo dunque i suoi stati mentali (ricordi, credenze, desideri, conoscenze). Secondo Baron-Cohen, questo è un segno incontrovertibile della necessità di una teoria della mente per spiegare cosa muove gli altri nelle loro azioni, dato che è molto più difficile generare spiegazioni non mentalistiche (ovvero non basate su stati mentali) per un insieme di comportamenti come questi, considerati molto semplici da interpretare<sup>88</sup>.

L'esperimento più diffuso per attestare e valutare le abilità di ToM messe in atto dagli individui è il cosiddetto *False Belief Task*<sup>89</sup> (Wimmer & Perner 1983), che, secondo ipotesi, permette di verificare se un individuo è capace di attribuire una falsa credenza, ovvero di capire se un soggetto può comprendere che gli altri possano rappresentarsi il mondo in un certo modo, e che questa rappresentazione può essere vera o falsa (Spaulding 2020). Lo scenario sperimentale prevede innanzitutto una storia, in cui sono presenti due personaggi: Sally e Anne. L'osservatore guarda Anne

---

<sup>88</sup> Un'interpretazione dello stesso evento in termini non mentalistici verrà fornita in §3.3.3, *pace* Baron-Cohen.

<sup>89</sup> La prima formulazione del *false belief task* è da attribuire al filosofo Daniel Dennett (1978), da cui Wimmer e Perner (1983) prendono ispirazione per la strutturazione del loro esperimento.

e Sally giocare con una bambola all'interno di una stanza; a un certo punto, Sally ripone la bambola con cui stava giocando all'interno di una scatola con il coperchio ed esce dalla stanza. Mentre Sally è fuori, Anne sposta la bambola dalla scatola in cui Sally l'aveva riposta a un'altra scatola. Sally poi rientra nella stanza per giocare di nuovo con la sua bambola. A questo punto, all'osservatore della scena viene posta una semplice domanda: dove cercherà la bambola Sally?

In un primo momento, la risposta potrebbe sembrare fin troppo scontata: Sally cercherà la bambola proprio dove l'aveva lasciata, nella prima scatola in cui lei stessa l'aveva riposta, visto che non può sapere che Anne ha spostato la bambola mentre era fuori dalla stanza. Riconosceremmo, dunque, che Sally ha una falsa credenza. Tuttavia, questa abilità interpretativa è proprio quella che si ritiene propria della ToM, e che l'esperimento in questione attesta come acquisita solamente tra i 4 e i 5 anni d'età (Frith & Frith 2005). Infatti, i bambini al di sotto di questa età non riescono ad assumere il punto di vista di Sally e ragionare con le sole informazioni che sanno lei possiede (e non le altre di cui loro sono a conoscenza, ovvero il fatto che Anne ha spostato la bambola). Solitamente, i bambini fino ai 4 anni rispondono a partire dalla loro posizione privilegiata di osservatore, dicendo che Sally cercherà la bambola nella scatola in cui l'aveva spostata Anne, migliorando poi nel tempo con l'acquisizione di una teoria della mente (Wellman et al. 2001). Oltre a investigare il processo di acquisizione di teoria della mente nello sviluppo dell'essere umano, l'esperimento di Sally e Anne ha anche messo in luce alcune situazioni in cui la teoria della mente è acquisita in fasi più tarde dello sviluppo, come nel caso dei soggetti affetti da disturbo dello spettro autistico (ASD, Baron-Cohen et al. 1985; Baron-Cohen 1995). In questi casi, ovvero se il soggetto osservatore è affetto da ASD, la capacità di attribuire le giuste credenze a Sally non è presente entro la soglia dei 5 anni di età (Happé e Frith 1995; Baron-Cohen 2000)<sup>90</sup>.

Grazie alla sua standardizzazione e all'importante ruolo che questo ricopre nella ToM, oggi il test di Sally e Anne è considerato essere una metodologia affidabile per la verifica di diversi meccanismi della teoria della mente (Byom & Mutlu 2013). In sostanza, l'esperimento serve a vedere se l'individuo soggetto dell'esperimento abbia acquisito la capacità di “leggere nella mente” dell'altro, capacità detta di *mindreading* (sinonimo di “teoria della mente”, anche detta “folk psychology” o “mentalizing”, cfr. Apperly 2011; Davies and Stone 1995; Goldman 2006; Nichols and Stich 2003). Tale capacità si basa sull'abilità del soggetto di i) rappresentare a livello concettuale gli stati mentali (credenze, desideri, percezioni) dell'altro; ii) connotare tali stati mentali di una intenzionalità, ovvero di concepire tali stati come “circa” qualcosa, “aventi una direzionalità” nel mondo; iii) comprendere in qualche modo la relazione che si instaura tra gli stati mentali dell'agente, le condizioni ambientali

---

<sup>90</sup> A partire da queste evidenze, il disturbo dello spettro autistico è stato caratterizzato soprattutto in quanto deficitario della capacità di teoria della mente (Happé e Frith 1995; Baron-Cohen 2000).

e il suo comportamento. In sostanza, il *mindreading* è alla base della capacità di elaborare una teoria della mente: se il soggetto dimostra di aver sviluppato tale capacità, significa che i) riesce a discretizzare gli stati mentali e comprenderne il significato (ad es. del concetto di credenza); ii) ha la capacità di attribuire stati mentali agli altri; iii) riesce a considerare l'altro come un agente che è mosso dai propri stati mentali, che a loro volta ne causano il comportamento.

Ora, la teoria della mente e le metodologie di verifica e sperimentazione a essa correlate sono state importate nel campo della HRI sia per quanto riguarda la validazione del robot come agente con capacità sociali, sia per fornire un modello su cui basarsi per costruire architetture cognitive che rendano capace il robot di comprendere le azioni e il comportamento del suo interlocutore umano. È utile fornire una panoramica di entrambe le questioni per comprendere a pieno il ruolo che la ToM assume negli studi di HRI.

### 3.3.1 *Le persone adottano la ToM per spiegare e prevedere il comportamento del robot?*

Capire se le persone adottano una teoria della mente per comprendere le azioni del robot significa verificare se il robot è considerato un agente sociale a tutti gli effetti. Secondo questa impostazione, se gli stessi meccanismi di ToM sono attivati per prevedere sia le azioni di un altro essere umano, sia quelle del robot, ciò implica che l'essere umano sta attribuendo degli stati mentali a entrambi, e giustifica il comportamento osservato come guidato da credenze, desideri e percezioni. In letteratura, tale questione ha preso il nome di *mind attribution* (e.g., Appel et al., 2020; Broadbent et al., 2013; Li and Sung, 2021; Lu et al., 2019; Stafford et al., 2014; Söderlund and Oikarinen, 2021; Wang and Krumhuber, 2018) e viene considerato un meccanismo fondamentale che prova l'effettiva considerazione del robot in quanto agente sociale. Il metodo più adottato per verificare la presenza di una teoria della mente nell'incontro con il robot consiste nella somministrazione di questionari diretti di natura verbale in cui si attesta la presenza di un vocabolario mentalistico, ovvero basato su termini e locuzioni che denotano considerazioni antropomorfe (come “credere”, “pensare”, ecc.) (Jost et al. (Eds.) 2020; Pekçetin et al. 2024), ma sono sempre più diffusi metodologie implicite di validazione come il *neuroimaging* (Hegel et al. 2008) e il tracciamento dello sguardo (Thellman et al. 2020). Numerosi studi forniscono evidenze che gli esseri umani adottano una teoria della mente per spiegare il comportamento di agenti non-umani, attribuendo loro quindi un certo livello di attività mentale, e ciò accade più facilmente se l'agente mostra qualità estetiche che ricordano l'essere umano (Gray et al., 2007; Gray and Wegner, 2012; Martini et al., 2016; Söderlund and Oikarinen, 2021; Hegel et al. 2008), elicitando reazioni simili a quelle in interazione con altri esseri umani (Epley, 2018; Foehr and Germelmann, 2020; Pelau et al., 2021; Reeves and Nass, 1996; Złotowski et al., 2015).

Tuttavia, in alcuni casi, le evidenze riportano una differenza tra l'osservazione di un agente umano e uno robotico nello stesso task. Ad esempio, replicando il *false belief task* in due condizioni (una condizione di controllo con individui umani, una condizione in cui Sally viene sostituita da una robot), Thellman et al. (2020) riscontrano una significativa differenza nelle risposte alla domanda: dove l'agente cercherà l'oggetto quando rientra nella stanza? I ricercatori sottolineano che, se per l'essere umano alla domanda viene data una risposta corretta nella quasi totalità dei casi, nella condizione del robot gli errori aumentano notevolmente e i soggetti sperimentati adducono motivazioni circa l'insicurezza del funzionamento della memoria del robot («“It depends if the robot have [sic] a memory of where it put the object”; “I don't think the robot remembers, i [sic] would be a lot to remember”; “It is unclear if the robot keeps a record of placed items”» Ibid.; 480), delle sue conoscenze o della sua possibilità di percepire gli oggetti nello stesso modo dell'umano («“I do not have enough information about what the program in the robot looks like for me to guess how it will behave”; “Unclear what sort of search algorithm or method the robot is running”», *Ibidem*). Tale ricerca sembrerebbe dimostrare, in linea con altre evidenze (Kiesler 2005), che la teoria della mente non venga adottata indistintamente nelle stesse circostanze quando l'agente è un robot o un umano.

In altri studi, Kahn et al. (2006, 2012) evidenziano che i bambini in età prescolare attribuiscono facilmente stati mentali a robot dalla forma animale o umanoide, ma si trovano in disaccordo con le proposizioni che descrivono il robot come “libero” o “con diritti”. Su questa base, in uno studio successivo (Kahn et al. 2011), i ricercatori consolidano l'ipotesi che i bambini in età prescolare non trattino i robot umanoidi nello stesso modo degli esseri umani, ma tendono a inquadrali in una categoria ontologica distinta, formulando la *New Ontological Category (NOC) hypothesis*.

In aggiunta alle discrepanze nell'attribuzione di stati mentali, diversi studi sottolineano le incongruenze dei risultati riportati per lo stesso setting sperimentale tra misure implicite e misure esplicite (Fiala et al. 2014; Thellman et al. 2022), notando che spesso se le misure implicite riscontrano l'attivazione di meccanismi di ToM di fronte al robot, le misurazioni esplicite vanno in direzione opposta.

In sostanza, da questi studi emerge il fatto che la verifica dell'attribuzione di stati mentali al robot, meccanismo fondamentale della ToM, non necessariamente comporta che il robot sia considerato un agente sociale al pari dell'essere umano. In ogni caso, e nonostante i risultati contrastanti, la ToM e i meccanismi di *mindreading* possono essere considerati come il quadro epistemologico di base attraverso cui l'interazione tra umani e robot viene indagata nei termini della cognizione sociale.



### 3.3.2 Come si può dotare il robot di una ToM?

Oltre ad essere uno strumento teorico pervasivo nell'indagine della postura umana in interazione con il robot, la ToM è considerata fondamentale da implementare nella architettura cognitiva robotica per fornire al robot la possibilità di comprendere al meglio le azioni dell'essere umano. Infatti, dal momento che è per lo più indiscusso il fatto che la cognizione sociale negli esseri umani funzioni con meccanismi di Teoria della mente, dotare il robot degli stessi meccanismi sembra essere l'unica soluzione per raggiungere un livello di intelligenza sociale complessa che possa interagire con l'essere umano. In questa prospettiva, gli studi pionieristici di Scassellati (2001, 2002) per dotare il robot di una teoria della mente sono stati seguiti da una nutrita schiera di esempi che tentano in modi diversi di implementare i meccanismi inferenziali tipici della ToM in una architettura adeguata. In particolare, Scassellati (2001) si concentra sulle possibilità di convertire le diverse versioni di ToM, elaborate rispettivamente da Leslie (2000) e Baron-Cohen (1995), nella costruzione di un'architettura cognitiva robotica con le caratteristiche individuate come fondamentali da entrambi gli approcci, che sarà poi implementata in Cog (Brooks et al. 1998; §2.3.2.2): *Pre-attentive Visual Routines* (relativa alla discretizzazione degli stimoli salienti, come il colore); l'attenzione visiva (che permette di rivolgere lo sguardo verso l'interlocutore o un oggetto di interesse comune, come nei meccanismi di *shared attention*); la discriminazione di un oggetto inanimato da uno animato; la capacità di seguire lo sguardo dell'agente (*gaze following*). Come detto, questi studi non sono stati che i primi di una lunga serie, che è in corso di sviluppo ancora oggi (Barchard et al., 2020; Sturgeon et al., 2019; Winfield, 2018; Bianco & Ognibene 2019a, 2019b; Ono et al. 2000). Riuscire a implementare le caratteristiche della ToM in un'architettura cognitiva viene considerato il tentativo più concreto di costruire robot che dimostrino delle capacità sociali complesse, seguendo anche esperimenti che dicono che più il robot dimostra capacità di *mindreading* di alto livello, più l'essere umano sarà propenso ad attribuirgli una mente, a trattarlo da agente e a costruire un rapporto più robusto in termini di fiducia (Mou et al. 2020; Esterwood & Robert 2023).

Date le premesse circa l'univocità dell'adozione di questa ipotesi per il funzionamento della cognizione sociale negli esseri umani, raggiungere l'obiettivo posto da questo filone di ricerche significherebbe, come auspicato da Scassellati, aprire un ampio spettro di possibilità per rendere il robot capace di meccanismi cognitivi complessi, che vanno dall'apprendimento alla comprensione dei propri stati mentali:

A robotic system that possessed a theory of mind would allow for social interactions between the robot and humans that have previously not been possible. The robot would be capable of learning from an observer using normal social signals in the same way that human infants learn; no specialized training of the observer would be necessary. The robot would also be capable of expressing its internal state (emotions, desires, goals, etc.) through social interactions without relying upon an artificial vocabulary. Further, a robot

that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly. (Scassellati 2001: 16)

Nei venti anni intercorsi tra i primi esempi di implementazione e l'attuale stato della ricerca, molti tentativi sono stati portati avanti, contribuendo di volta in volta a sviluppare aspetti specifici della ToM per le architetture robotiche, ma senza arrivare a una vera e propria svolta in termini di qualità dell'interazione. Infatti, il processo di implementazione delle capacità di comprensione delle intenzioni dell'essere umano da parte del robot sembra incontrare una serie di problemi relativi al modo di implementare tale teoria nell'architettura cognitiva (Kim & Lipson 2009; Bianco & Ognibene 2019a) e, come vedremo (§3.3.4), sembrano comportare una serie di problemi di difficile risoluzione. In più, anche se questa serie di problemi dovesse essere risolta, alcuni studi dimostrano che anche se il robot esibisce capacità di teoria della mente in dialogo con l'essere umano, ciò non comporta automaticamente un maggiore grado di accettabilità e non ha effetti diretti sullo stato dell'interlocutore umano (Benninghoff et al. 2013) o sull'aumento della fiducia percepita (Romeo et al. 2022).

Tuttavia, la scommessa di riuscire a costruire architetture cognitive che si basino sulla ToM e che riescano a far emergere una capacità sociale complessa non ha ancora esaurito la sua spinta.

### 3.3.3 *Not the only game in town*

Lo stato dell'arte che abbiamo delineato fino a qui rende l'idea della centralità della ToM sia nel campo delle scienze cognitive, sia negli studi di HRI. Di fatto, la ToM è considerata la teoria a cui fare riferimento universalmente per spiegare i meccanismi di comprensione che avvengono nella cognizione sociale, venendo importati in maniera non problematica anche all'interno della HRI.

Il proposito che Dautenhahn (2007b) delinea riguardo la necessità di impostare il progetto di realizzare agenti capaci di interagire in modo efficace sulla base della natura sociale dell'intelligenza umana, quindi, si trasforma nel tentativo di implementare i meccanismi di ToM nelle architetture cognitive, da una parte, e dall'altra nell'applicazione del *framework* teorico del *mindreading* e della *folk psychology* per comprendere quale sia la reazione dell'essere umano di fronte al robot. I due percorsi di ricerca che seguono i precetti della ToM trovano perfetta compatibilità con la sfida lanciata da Dautenhahn, dal momento che la teoria che lei propone come riferimento per spiegare la natura sociale dell'intelligenza umana, la *social intelligence hypothesis*, è essa stessa sviluppata sulla base della teoria della mente. Infatti, secondo Dunbar (1998), l'evoluzione di un cervello sociale nei primati superiori è direttamente correlata alla necessità di attribuire stati mentali agli altri in un sistema abitato da più individui della stessa specie, attività che ha comportato la specializzazione di

parte del cervello nel compito di *mindreading* prima e poi, solamente sulla base di queste capacità inferenziali, la comparsa di un'intelligenza non-sociale, come quella mostrata nel *problem-solving* e nella navigazione dell'ambiente.

Ora, anche sulla scorta di ciò che abbiamo osservato nel capitolo precedente circa l'evoluzione dei concetti di cognizione e di intelligenza nella rivoluzione della *Nouvelle AI*, possiamo elaborare alcune riflessioni circa il rapporto che intercorre tra i) il proposito di fondare l'implementazione dell'intelligenza generale su quella sociale; ii) la natura inferenziale della ToM; iii) il fatto di adottare la ToM come meccanismo universale di cognizione sociale.

i) *Intelligenza sociale e intelligenza generale*

Abbiamo visto nel capitolo precedente come l'impostazione della AI classica e delle scienze cognitive computazionaliste e funzionaliste si basasse sulla definizione dell'intelligenza umana a partire dal presupposto che ciò che la costituiva era il pensiero logico-proposizionale, elaborando un concetto di intelligenza generale<sup>91</sup> propria solamente della specie umana interamente costruita sul principio logico-proposizionale e simbolico. In questo processo, anche i meccanismi e le azioni non immediatamente riconducibili a – o identificabili primariamente come – meccanismi logico-proposizionali sono stati riportati a questo paradigma, trasformando ogni attività portata avanti dall'essere umano in un processo di risoluzione di problemi e ragionamento per euristiche (§2.1). Le attività di visione, di navigazione, di manipolazione di oggetti sono state trattate modellizzandole attraverso processi inferenziali, definendo ogni passaggio come un'attività di recupero o manipolazione di simboli. Come si è visto, per l'intelligenza sociale non è accaduto diversamente: nel quadro della ToM, anche la comprensione degli altri funziona come un meccanismo di risoluzione di problemi e ragionamento per euristiche, caratterizzato dalla capacità – a volte considerata completamente innata, a volte appresa nelle prime fasi dello sviluppo a partire da capacità innate<sup>92</sup> – di attribuire stati mentali agli altri. Anzi, secondo Dunbar – e con lui Dautenhahn – è proprio dal fatto che la cognizione sociale e l'intelligenza che la permette funzionano così che è stato possibile per la specie umana sviluppare il *problem-solving* e altre attività proprie dell'intelligenza

---

<sup>91</sup> Tali considerazioni sono quelle che ancora sorreggono l'impresa, ad oggi tornata in primo piano con la tecnologia della AI generativa (Jo 2023), di costruire una *Artificial General Intelligence* (Goertzel et al. 2014), ovvero un'intelligenza artificiale che eguagli quella umana in tutti i suoi aspetti.

<sup>92</sup> Il dibattito tra coloro che sostengono che la ToM ha una fortissima componente innata (Fodor 1983; Baron-Cohen 1995; Leslie et al. 2004) e quelli che argomentano a favore di una sua comparsa nel corso dello sviluppo ontogenetico del bambino (Gopnick & Meltzoff 1997) è ancora oggi molto acceso, e anzi costituisce uno dei punti più discussi da parte dei teorici della mente. Nonostante la sua importanza, la sua esposizione e trattazione va oltre le necessità esplicative di questa tesi. Per una panoramica sulle diverse posizioni, che a volte possono risultare anche non così lontane l'una dall'altra, si veda Leslie (2000); Meltzoff (2011).

umana, considerate capacità non-sociali. La ToM, elaborata proprio a partire da un'idea di cognizione basata su meccanismi inferenziali di ordine logico e formata da stati mentali di natura proposizionale, ha fornito alla HRI un appiglio comodo e adeguato all'interno delle scienze cognitive visto che adottano entrambe una concezione della mente come quasi esclusivamente analizzabile in termini simbolici e proposizionali, dominata ancora dalla metafora della mente come computer.

Questa operazione, però, è basata su un cortocircuito. Infatti, a ben vedere, più che basare l'intelligenza generale umana su quella sociale, sembra che l'operazione sia esattamente quella contraria: nella visione della ToM, è proprio il meccanismo centrale all'intelligenza generale, quello del ragionamento logico-proposizionale, che sta alla base delle possibilità di un'intelligenza sociale, in cui tale ragionamento è esercitato nella spiegazione del comportamento degli altri tramite inferenza e attribuzione di stati mentali. L'argomentazione, allora, diventa circolare, e scegliere su quale delle due ipotesi basare l'altra diventa una questione di interessi e vantaggi. Perché dire che la cognizione sociale sta alla base della intelligenza generale se il meccanismo inferenziale che ne consente l'evoluzione è quello che contraddistingue l'intelligenza umana generale, considerata una modalità cognitiva non-sociale? Per di più, se guardiamo alla storia delle applicazioni dell'AI, notiamo che i tentativi volti a replicare il pensiero di alto livello basato sui principi dell'intelligenza logico-proposizionale (come dimostratori di teoremi, programmi che simulano giocatori di giochi da tavolo, elaboratori di linguaggio come i più recenti programmi di AI generativa) hanno avuto risultati di gran lunga migliori rispetto a quelli raggiunti con l'intenzione di modellizzare l'intelligenza sociale. Ciò è probabilmente dovuto al fatto che, se per i meccanismi logico-proposizionali che la GOF AI ha creduto la chiave della capacità di pensiero umano abbiamo descrizioni e formalizzazioni molto approfondite, con definizioni che stabiliscono i confini tra i concetti e una metodologia basata sulla logica atta a implementarli, per la cognizione sociale il percorso di definizione e comprensione è ancora oggi in corso. Infatti, non c'è accordo nello stato dell'arte di tutte le discipline che abbiamo chiamato in causa fino ad ora che esprima un largo consenso né su cosa la cognizione sociale sia, né sulle sue modalità di realizzazione nelle dinamiche interazionali, anche all'interno del dibattito sulla ToM. Torna, in questo caso, il monito di Moravec con il suo paradosso, il quale ha cercato di puntare l'attenzione sulla difficoltà di modellizzare processi cognitivi all'apparenza meno complessi, ma che si rivelano essere i più ostici da sottoporre all'impresa: la cognizione sociale sembra essere proprio uno di questi casi.

Ciò che stiamo mettendo in discussione qui non è la validità dell'ipotesi di ricondurre lo sviluppo filogenetico dell'intelligenza umana per come si mostra oggi a una fondazione nel dominio della cognizione sociale. Piuttosto, ciò che vogliamo contestare è l'idea di basare l'attività di replicare i meccanismi di cognizione generale umana seguendo il modello derivante da *questa particolare* concezione e teorizzazione di intelligenza sociale. L'ipotesi di fondare le capacità dell'intelligenza umana su quella sociale ha subito incredibili sviluppi nel corso degli anni, sia nel campo delle neuroscienze (Ferrari e Gallese 2007; Rizzolatti et al. 2002), sia nelle scienze cognitive, in psicologia e in filosofia (Fuchs & De Jaegher 2009; Gallagher 2013; Lindblom 2015; Paolucci 2021; Tomasello 1998), ed è stata elaborata ulteriormente soprattutto a seguito dell'*intersubjective turn* (De Jaegher 2018) e all'interno delle scienze cognitive *embodied* ed *enattiviste* (Morganti et al. 2008). Quella che si delinea è, allora, la possibilità di sciogliere il cortocircuito tra intelligenza generale e intelligenza sociale guardando agli sviluppi recenti riguardo alle modalità esplicative e descrittive delle capacità sociali dell'essere umano.

ii) *La natura inferenziale della ToM*

Il motivo per cui si crea il cortocircuito appena esposto tra intelligenza generale e intelligenza sociale è che entrambe sono basate sullo stesso meccanismo di fondo, radicato come detto nella concezione logica-proposizionale della cognizione: l'inferenza. Nell'ottica della ToM questo accade perché, come fatto notare in più occasioni da vari interpreti (Gallagher 2005, 2008, 2012; 2020; Hutto 2008; Gallagher & Hutto 2008, Froese & Gallagher 2012), la mente è localizzata all'interno dell'individuo, nascosta ai suoi interlocutori, che di fatto la guardano dall'esterno, spettatori di ciò che accade. Tale concezione fortemente internalista implica che per comprendere effettivamente *cosa passa per la testa dell'altro* si debba decrittare il rapporto logico che lega il comportamento esibito con gli stati mentali che l'hanno causato, elaborando inferenze circa le cause (credenze, desideri, intenzioni, percezioni) nascoste nella mente e inaccessibili che portano a quel comportamento. Un rapporto di causazione, quindi, che deve essere indovinato a partire dagli indizi che l'agente semina, tradendo auspicabilmente la natura dei suoi stati mentali, attraverso la formulazione di una teoria<sup>93</sup>. Ora, porre al centro della capacità di comprendere gli altri il meccanismo inferenziale significa continuare sulla via battuta dalla GOFAI, che prevedeva la formalizzazione in

---

<sup>93</sup> Non stupisce, dunque, che alcuni propositori della ToM, concepita come elaborazione di una vera e propria teoria, vedano delle somiglianze tra il processo di acquisizione e formazione della teoria nelle fasi di sviluppo del bambino e il processo di strutturazione di una teoria scientifica. Cfr. Gopnik (1996).

termini logici di tutte le attività cognitive dell'agente umano. I problemi che ha comportato l'adozione di questa visione, ad esempio in SHAKEY (§2.1.2), erano sicuramente dati anche dallo stadio embrionale delle tecnologie impiegate per la sua costruzione, ma il cambiamento di prospettiva operato dalla *Nouvelle AI* ci ha mostrato come l'implementazione del comportamento sulla base di principi diversi dall'inferenza massiccia e pervasiva fosse in larga misura più efficace per consentire al robot di interagire con il suo ambiente. Ciò che questa storia può ulteriormente mostrarci, allora, è che l'implementazione di processi sensomotori può fornire una via d'uscita all'*empasse* creata dalle condizioni generate dalla concezione inferenziale della cognizione per ogni sua attività, che implica la necessità di trovare una soluzione in termini di capacità computazionale legata alla rappresentazione dell'ambiente. In sostanza, l'operazione consiste nello smettere di guardare alla cognizione sociale esclusivamente attraverso lo strumento dell'inferenza logica, ma immergersi nello studio del modo in cui l'agente si coordina al suo ambiente in interazione, soprattutto nella sua componente corporea e situata. Come abbiamo osservato già nella ricostruzione fornita nel capitolo precedente, la componente interazionale, *embodied* e situata dell'essere umano è già stata messa in luce da diversi studi a partire dalla fine degli anni '70 (§2.3) e propone di considerare la cognizione e la mente umana guardando ai meccanismi sensomotori di accoppiamento fra l'organismo e l'ambiente.

La stessa ricercatrice canadese, in diversi articoli (Dautenhahn 2007a, 2014; Fong et al. 2003) afferma che le evidenze provenienti dalla robotica testimoniano la necessità di uno spostamento di focus verso gli accoppiamenti sensomotori tra organismo e ambiente. In particolare, nello stesso articolo in cui propone di abbracciare la *social intelligence hypothesis*, argomenta:

While progress has been made in the domain of what is considered now 'classical AI', e.g. chess-playing programs are able to beat expert human chess players, [...] from the perspective of trying to understand or create human intelligence, it has become apparent that such skills are not necessarily those that 'make us human'. Also, attempts to put 'AI' on wheels, i.e. to design AI robots, illuminated a fundamental problem with the view of intelligence as 'disembodied' and 'symbolic', i.e. getting a robot to do even very 'simple things' [...] turned out to be very surprisingly difficult. [...] More fundamentally, skills involving sensing and acting and close couplings between these in order to deal with the dynamics and unpredictability of the 'real world' have become the new big challenges. [...] More recently, sensorimotor skills emphasizing the embodied nature of human intelligence (including locomotion, object manipulation, etc.) are considered to be the more fundamental but certainly more biologically and developmentally plausible milestones that researchers are aiming at, highlighting the close relationship between mind, body and environment [...]. (Dautenhahn 2007b: 679-680).

Dunque, se le capacità sensomotorie sono considerate essere più fondamentali delle abilità di cognizione simbolica e proposizionale, allora la ToM non risulta essere la teoria di cognizione sociale più adatta, dal momento che non può renderne conto in modo diverso da come concettualizzato all'interno della teoria computazionale della mente. Concentrare l'attenzione sulla cognizione sociale e sul modo in cui l'essere umano instaura regimi di significazione interazionali significa adottare una teoria di cognizione sociale che ponga alla base i processi sensomotori e non i meccanismi inferenziali tipici della ToM.

Per sorpassare la visione imposta da questo tipo di teorie, quindi, dobbiamo adottare una visione della cognizione sociale che riconosca il ruolo costitutivo del corpo in interazione, del contributo che ogni agente apporta nella costruzione dell'interazione e che concepisca il processo di comprensione dell'altro non in termini di ragionamento inferenziale. Di conseguenza, ciò implica anche abbandonare una visione della mente dell'altro interamente confinata all'interno dell'individuo, per decifrare le intenzioni del quale c'è bisogno di ricorrere a teorie a loro volte interne o a meccanismi cablati nel cervello.

iii) *Teoria della mente come modello universale di cognizione sociale*

Come è possibile, quindi, rendere conto dei meccanismi che strutturano la nostra comprensione degli altri se non possiamo fare affidamento sulla teoria più diffusa all'interno delle scienze cognitive?

Adottare la ToM come teoria universale per spiegare la cognizione sociale come se fosse l'unica opzione valida non restituisce giustizia all'acceso dibattito tutt'ora in corso all'interno delle scienze cognitive. Infatti, lungi dall'essere “the only game in town”<sup>94</sup> (Fodor 1975) come potrebbe apparire guardando agli studi in HRI, negli ultimi anni la ToM è stata oggetto di numerose riflessioni critiche, che hanno motivato la formulazione di teorie alternative per comprendere la cognizione sociale e i suoi meccanismi. Soprattutto a partire dalle riflessioni sviluppate nei punti precedenti, allora, e vista la necessità di rendere conto dei processi basilari della cognizione sociale in termini corporei e situati, è utile rivolgere la nostra attenzione alle teorie della cognizione sociale sviluppate dalla corrente enattivista delle scienze cognitive, così da poter finalmente basare la nostra comprensione di ciò che accade nell'interazione tra umani e robot su una dimensione incarnata e sensorimotoria. Infatti, in opposizione a un approccio alla cognizione sociale

---

<sup>94</sup> L'espressione “the only game in town” è diventata celebre dopo che Fodor (1975) l'ha impiegata in difesa dell'ipotesi del linguaggio del pensiero (*Language of Thought*, LoT). Secondo il filosofo, l'ipotesi del LoT si prospettava (almeno in quel periodo) come unica soluzione valida a cui aderire per cercare di spiegare la cognizione, intesa ovviamente come computazione e impossibile da concepire senza rappresentazioni (“no cognition without representation” è un altro slogan coniato all'interno della stessa opera).

in terza persona (spettatoriale), disincarnato e basato sul *mindreading*, le più recenti formulazioni enattiviste (Hutto 2008; Hutto et al. 2011; Reddy & Morris 2004; Ratcliffe 2007; Gallagher 2017, 2021; Gallagher & Zahavi 2008; Fuchs & De Jaegher 2009; De Jaegher & Di Paolo 2007; de Bruin & de Haan 2012) propongono una visione in prima e seconda persona (in cui cioè l'agente è coinvolto attivamente nell'interazione), basata sulle pratiche corporee e che coinvolge diverse capacità dell'individuo – l'imitazione, l'attribuzione dell'intenzionalità, il *gaze following*, ecc. – che non dipendono esclusivamente da meccanismi inferenziali di comprensione come il *mindreading*, ma da esperienze incarnate vissute dall'individuo in quanto organismo in stretto accoppiamento con il proprio ambiente fin dalla nascita<sup>95</sup>.

Per mostrare la differenza tra gli approcci non solamente in termini di presupposti teorici, ma anche guardando agli effetti sulla considerazione dei fenomeni, possiamo vedere come l'esperimento mentale proposto da Baron-Cohen (1995) ed esposto in precedenza possa essere riletto senza troppi sforzi attraverso una spiegazione non necessariamente mentalistica. Infatti, in questa ottica, trovandosi nella situazione in cui qualcuno entra nella stanza in cui siamo, si guarda intorno ed esce senza dare spiegazioni, l'interpretazione in termini mentalistici di questa sequenza di azioni viene resa possibile da una più profonda e radicata esperienza corporea, costruita sulla base di una storia personale di interazioni simili che permette di comprendere l'altro innanzitutto in termini di *azioni incarnate* (Varela et al. 1991; Gallagher 2005; Hutto 2008; De Jaegher & Di Paolo 2007). Non c'è bisogno di spiegare il comportamento della persona ricorrendo a inferenze mentalistiche che ci facciano dedurre cosa la muova per capire lo strano comportamento dell'individuo di fronte a noi. Capiremmo immediatamente guardando ai suoi movimenti corporei, alle sue espressioni facciali, alla direzione del suo sguardo cosa la persona sta cercando di fare, e questo perché abbiamo – in quanto esseri umani cresciuti in un ambiente sociale – esperienza di cosa significhi essere in una stanza per cercare qualcosa. Così, la mente non è considerata come qualcosa di interno all'individuo, ma è percepibile sulla superficie delle azioni dell'altro:

« [...] in seeing the actions and expressive movements of the other person in the context of the surrounding world, one already sees their meaning: no inference to a hidden set of mental states [...] is necessary. When I see the other's action or gesture, I see [...] the meaning in the action or gesture. I see the joy or I see the anger, or I see the intention in the face or in the posture or in the gesture or action of the other» (Gallagher 2008: 542)

---

<sup>95</sup> Un interessante filone di studi anticipa il momento in cui l'organismo si allaccia inscindibilmente al suo ambiente al processo di formazione del feto nell'utero della madre. Cfr. Ciaunica et al. (2021a, 2021b).



La situazione che ci troviamo ad affrontare delinea un quadro disciplinare complesso, in cui il contributo che le scienze cognitive hanno dato fino ad ora per chiarire le varie dimensioni della HRI ha comportato un cortocircuito esplicativo. Il percorso che stiamo compiendo, dunque, si definisce proprio rinunciando alla ToM come unica teoria della cognizione sociale, adottando altresì un approccio che possa rendere conto dei processi fondamentali di interazione a livello sensorimotorio e incarnato.

Per concretizzare il cambiamento di sguardo che abbiamo proposto, il primo passo consiste nel mostrare come il contributo della teoria enattivista della cognizione sociale possa rendere chiare le modalità in cui gli agenti costruiscono regimi interazionali atti alla comprensione reciproca non attraverso percorsi inferenziali logico-proposizionali, ma sulla base di una serie di *azioni incarnate*. Ciò ci consentirà di apportare alcuni cambiamenti alla concettualizzazione dell'interazione tra umani e robot.

#### 3.3.4 Enattivismo e cognizione sociale: dall'inferenza all'interazione

Fino a questo momento abbiamo parlato della Teoria della Mente senza specificare le diverse varianti che ne costituiscono la dialettica interna. Il motivo per cui abbiamo trattato la ToM come un'unica ipotesi esplicativa è che, sebbene ognuna delle sue varianti mostri una propria identità rintracciabile nelle specificazioni date al processo di comprensione intersoggettiva e nella diversa considerazione di alcuni fattori per il suo raggiungimento, queste condividono il nucleo centrale della loro struttura esplicativa: l'attribuzione di stati mentali e il *mindreading*.

In questo paragrafo, cercheremo di rendere in dettaglio la specificità delle varianti della ToM poiché funzionale a rendere conto della validità della teoria enattivista della cognizione sociale non solo ed esclusivamente per meccanismi inferenziali di natura logico-proposizionale molto vicini a una tradizione cognitivista e funzionalista, ma anche per le varianti che considerano in qualche modo il contributo del corpo (e, nello specifico, del cervello) per l'attivazione dei meccanismi di comprensione dell'altro. Infatti, nonostante il ruolo del corpo, almeno negli ultimi 30 anni, sia stato considerato sempre più all'interno dell'equazione cognitiva – a partire dalla svolta *embodied* della cognizione (Clark 1999; Ceruti & Damiano 2013; Chemero 2009; Shapiro & Spaulding 2024) – ciò è avvenuto in modi diversi. Da una parte, è possibile individuare una serie di teorie e approcci delle scienze cognitive definite come *weak* o *simple embodiment* (cfr. Chemero 2009; Gallagher 2011), in cui il corpo è visto come il supporto sul quale la mente si realizza, che deve a sua volta essere integrato (spesso rappresentazionalmente) all'interno dell'architettura cognitiva, la cui definizione ancora converge con quella data dalla scienza cognitiva computazionalista. In questo senso, le teorie

dell'*embodiment* debole non fanno altro che introdurre all'interno dell'equazione computazionale della mente e della cognizione una nuova variabile, il corpo, che deve essere rappresentato dalla mente come ogni altro elemento al di fuori di essa, e che non svolge un ruolo costitutivo nell'emersione delle capacità cognitive. Al contrario, *l'approccio radicale alla cognizione*, di cui quello enattivista è una variante (Chemero 2009; Ceruti e Damiano 2013), considera il corpo come un elemento costitutivo nella creazione della possibilità di una mente, sullo stesso piano dell'ambiente e del cervello dell'organismo, adottando una visione autopoietica e dinamica del vivente (Maturana e Varela 1980; Varela et al. 1991; Damiano 2010). Tale divisione è cruciale per capire perché, anche se il ruolo del corpo viene incluso all'interno di alcuni approcci di ToM, questo non viene considerato come elemento costitutivo, né imprescindibile, per la costituzione di alcun tipo di cognizione.

Dunque, come sottolineato in più sedi (Spaulding 2020; Gallagher 2005, 2007, 2017, 2020; Gallagher & Hutto 2008; Hutto 2008), di pari passo con lo sviluppo di approcci deboli e radicali alla cognizione, il dibattito riguardo il funzionamento e i meccanismi che rendono possibile la costruzione o l'evoluzione di una ToM è stato dominato da due alternative: la *Theory theory* e la *Simulation Theory*.

Rispettivamente, la *Theory theory* (d'ora in avanti anche TT) declina la ToM attraverso l'utilizzo delle teorie della *folk psychology*. Secondo questa variante, le inferenze attraverso cui comprendiamo gli altri sono rese possibili da un insieme di teorie "del senso comune" o *folkpsicologiche*, tramite cui costruiamo la nostra teoria riguardo i probabili stati mentali che ne causano il comportamento. Queste inferenze vengono guidate dai principi psicologici che nella *folk psychology* connettono gli stati mentali alle azioni dell'altro: da questi nessi deriverebbero quindi predizioni e spiegazioni del comportamento degli altri (Carruthers & Smith 1996; Gopnick & Wellman 1992; Nichols & Stich 2003). Nella sua forma più generale, la TT afferma che il *mindreading* è un processo che necessita di una grande quantità informazione (*information-rich*): il processo di comprensione degli stati mentali dell'altro ha bisogno di e viene permesso da un'ingente quantità di dati psicologici. Negli anni, sono state sviluppate diverse versioni della TT, soprattutto in riferimento alla natura innata (Baron-Cohen 1995; Leslie 1992) o meno (Gopnick e Meltzoff 1997) della capacità di comprendere le altre menti. In questo account, il ruolo del corpo è assolutamente irrilevante, considerato al più come strumento percettivo attraverso cui vengono raccolte le sensazioni e supporto alle elaborazioni cognitive di natura inferenziale.

D'altra parte, la *Simulation Theory* (d'ora in avanti anche ST) rifiuta la possibilità di inferire direttamente gli stati mentali altrui, ipotizzando invece che l'individuo spieghi e predica il comportamento dell'altro impiegando i propri stati mentali come modello, mettendosi di fatto nei panni dell'altro e immaginando quali sarebbero i propri stati mentali se si trovasse nella stessa

situazione che sta cercando di spiegare e quale sarebbe il comportamento che ne deriverebbe. In particolare, la ST ipotizza che la cognizione sociale funzioni grazie a una simulazione interna: l'individuo ha una propria mente, che utilizza come modello per simulare gli stati mentali dell'altro e, in seguito, gli attribuisce le conclusioni che derivano dal suo ragionamento su quelli che ipotizza essere i suoi stati mentali (Currie & Ravenscroft 2002; Goldman 2006; Gordon & Cruz 2003; Nichols & Stich 2003). Negli ultimi anni, la scoperta dei neuroni specchio (Rizzolatti et al. 1999) è stata interpretata da alcuni come evidenza a favore della ST, individuando nel meccanismo di attivazione di questo insieme di neuroni sia quando si sta svolgendo un'azione, sia quando si guarda qualcun altro svolgere la stessa azione, come la sede cerebrale in cui la simulazione avviene (Gallese & Goldman 1998). In questa versione, il ruolo del corpo è più importante rispetto alla TT, ma solamente in quanto base neurale che permette la realizzazione della simulazione. Ciò alleggerisce la necessità di ricorrere a meccanismi esplicativi in forma informazionale esplicita. Infatti, al contrario della TT, la ST non necessita di una ingente quantità di informazioni per entrare in funzione (*information-poor*): la comprensione degli altri non ha bisogno di un bagaglio di conoscenze in forma proposizionale molto grande, né tanto meno che queste siano in forma di teorie folkpsicologiche; ciò di cui ha bisogno l'individuo è solamente la capacità di afferrare il significato di cosa l'altro stia facendo in una data situazione e attribuirgli le conclusioni derivate dalla simulazione.

In sostanza, sebbene la TT e la ST siano in disaccordo su come avvenga l'attribuzione di stati mentali all'altro, concordano sul fatto che i) il modo in cui comprendiamo gli altri nei contesti sociali sia tramite la spiegazione e la previsione del loro comportamento sulla base di attribuzione di stati mentali, ovvero sulla formazione di una teoria della mente, e ii) che la spiegazione riguardo alle modalità di comprensione e interazione con gli altri sia ciò che abbiamo già definito *mindreading*. In questo, entrambe le varianti della ToM restano ancorate a una visione della mente *nuda* (Ceruti & Damiano 2013), in cui la cognizione non è assolutamente dipendente dall'azione pragmatica sviluppata corporalmente in un ambiente, ma ancora il risultato del funzionamento dei meccanismi cognitivi disincarnati e realizzata nella struttura simbolica.

Raccogliendo questi elementi teorici, Gallagher (2021: 71) identifica tre caratteristiche su cui entrambe le alternative basano le proprie ipotesi sul funzionamento della cognizione sociale:

- *The unobservability principle*: la TT e la ST presuppongono che la cognizione sociale sia un problema dovuto al fatto che non si possa avere accesso diretto agli stati mentali dell'altro. Chiamato anche il “problema delle altre menti” (Avramides 2023), il punto secondo queste teorie è che non possiamo percepire direttamente i pensieri, le sensazioni o le intenzioni dell'altro, ma abbiamo bisogno di un processo cognitivo extra-percettivo, ovvero il *mindreading*, per inferire o simulare quali possano essere i suoi stati mentali.

- *The observational stance*: in entrambe le versioni, l'atteggiamento standard verso gli altri è di tipo osservazionale, ovvero in terza persona. Infatti, sia nella TT che nella ST, l'osservazione dell'altro è fondamentale per poi elaborare una spiegazione in termini mentalistici o riuscire a prevedere ciò che accadrà dopo.
- *The supposition of universality*: il solo e unico modo attraverso cui comprendiamo gli altri è l'inferenza basata sulla *folkpsychology* (TT) o la simulazione (ST)<sup>96</sup>.

A partire da queste considerazioni, Gallagher individua una serie di problematiche derivanti dalla adozione di queste teorie per la cognizione sociale (Ibid.: 76-97), come il fatto che queste teorie non spieghino come sia possibile individuare i tratti rilevanti di un comportamento per poterlo considerare un indizio degli stati mentali dell'altro (*starting problem*), o che l'attivazione di un determinato pattern neuronale di per sé non comporta direttamente la simulazione degli stati mentali dell'altro, quando più una loro percezione (*matching problem*). In sostanza, per come elaborato all'interno di ognuna delle due varianti della ToM, Gallagher rileva che la comprensione dell'altro passa prima di tutto tramite una percezione enattiva delle sue azioni in interazione.

Dunque, l'elaborazione di una teoria della cognizione sociale basata sulla rilevanza e la significatività dell'interazione emerge come una necessità esplicativa. L'*Interaction Theory* (IT, Froese & Gallagher 2012; Gallagher 2001, 2008a, 2008b, 2020; Gallagher & Hutto 2008) si configura come il tentativo di conferire preminenza all'interazione viva a partire dalla considerazione delle *embodied actions*, ovvero *azioni incarnate* degli individui che si trovano a interagire. Tale operazione risulta cruciale per ancorare la comprensione della cognizione sociale alla dimensione corporea, dinamica e sensorimotoria, come suggerito da Varela, Thompson e Rosch nel testo cardine della teoria enattivista della cognizione:

By using the term *embodied* we mean to highlight two points: first that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context. By using the term *action* we mean to emphasize once again that sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition. (Varela et al. 1991: 172-173)

---

<sup>96</sup> Date le ampie convergenze, negli ultimi anni, a seguito di un dibattito durato a lungo su ciò che definiva le differenze di ciascun approccio, sono emersi alcuni tentativi di integrazione. Tra le versioni integrate più diffuse, vi è quella che individua le capacità di lettura e simulazione interna dell'altro richiamate dalla ST come base esperienziale su cui le capacità folkpsicologiche proprie della TT possono essere impiegate per elaborare spiegazioni sul comportamento dell'altro (Apperly 2012; Spaulding 2018). Gallagher (2020: 72) identifica una doppia modalità di integrazione delle diverse posizioni: una teoria *ibrida* è una teoria in cui una delle due versioni è posta come meccanismo di base, e l'altra è chiamata in suo supporto in casi particolari; d'altra parte, un approccio *pluralista* concepisce l'uso alternato delle due versioni a seconda della circostanza.

Nel dominio della cognizione sociale, tale precetto si configura come la fondazione di un processo in cui entrambi gli individui coinvolti costruiscono comprensione prima di tutto sulla base di una dimensione incarnata della cognizione, influenzata dal contesto biologico, psicologico e culturale che costituisce l'ambiente all'interno del quale questa si dispiega. Infatti, secondo Gallagher (2001: 90), «[b]efore we are in a position to form a theory about or to simulate what the other person believes or desires, we already have specific pre-theoretical knowledge about how people behave in particular contexts».

Nell'elaborazione della IT, l'acquisizione di questa conoscenza pre-teorica che sta alla base della possibilità di comprendere l'altro in interazione è acquisita nel periodo di intersoggettività primaria (Trevarthen 1979), ovvero in quel periodo dello sviluppo ontogenetico del bambino che va dalla nascita ai 9 mesi di vita. È in questa fase, che si basa su capacità sensorimotorie innate o sviluppate molto presto dopo la nascita, che gli individui entrano in relazione con gli altri (Reddy 2008; Reddy & Morris 2004) e, a partire da queste capacità, costruiscono un *saper-fare (know-how)* riguardo la cognizione sociale, una comprensione dell'altro essenzialmente pragmatica (Gallagher 2021: 105) che riguarda le reali o potenziali azioni che l'agente può intraprendere con il suo interlocutore. Una serie di *affordance sociali*, dunque, che vengono direttamente percepite attraverso abilità che non necessitano di arrivare a un livello linguistico o proposizionale: «We understand the actions of others in terms of the affordances that they, and the situation, offer» (Ibid.: 107). È importante sottolineare che tali capacità interpretative basilari di natura sensorimotoria sono sì primarie perché sviluppate per prime, giacché emergono in quella fase della vita in cui alcun altro mezzo è dato all'individuo per costruire comprensione, ma non vengono abbandonate una volta che la fase termina: sono primarie anche perché hanno preminenza nel quotidiano e regolare modo di comprendere gli altri sempre, in ogni momento della vita. La percezione e comprensione dell'altro in interazione continua quindi a basarsi prima di tutto sull'intersoggettività primaria, intesa come quell'insieme di capacità sensomotorie che sviluppiamo ed evolviamo interagendo direttamente e a livello corporeo con gli altri in un ambiente sociale, anche quando attraversiamo fasi di intersoggettività successive<sup>97</sup>. Ciò è dimostrato da diversi studi che testimoniano, ad esempio, che le espressioni facciali, i gesti e le azioni che avvengono in contesti quotidiani vengono interpretate anche dagli adulti, e non solo dai bambini,

---

<sup>97</sup> In psicologia dello sviluppo, oltre alla intersoggettività primaria (0-9 mesi, che si instaura soprattutto nella relazione diadica *caregiver*-bambino), è individuata quella secondaria (9-12 mesi), in cui il bambino impara a comprendere le intenzioni dell'altro anche in relazione a un oggetto terzo (capacità detta di *joint attention*), avviando la transizione dalla relazione a tema alla relazione integrata (Trevarthen 1998), e quella terziaria (12-18 mesi), in cui il bambino impara a costruire mondi possibili e a navigarli (*as-if function*, Fuchs 2017) elaborando pensieri complessi circa le azioni reali o possibili degli altri (Violi 2012; Paolucci 2021). È in questa ultima fase che il bambino comincia ad addentrarsi nel sistema del linguaggio e a sviluppare una consapevolezza di sé e del proprio corpo, come testimoniato dal raggiungimento del riconoscimento allo specchio (Lobaccaro & Bacaro 2022).

attraverso abilità corporee interazionali che vengono esercitate in tempo reale (Lindblom 2015; Lindblom & Ziemke 2008)<sup>98</sup>.

L'IT, allora, viene definita proprio riconoscendo l'importanza pervasiva delle capacità sensorimotorie considerate primarie nell'interazione, nel doppio senso articolato poc'anzi:

IT contends that understanding other people is primarily based on [...] forms of embodied practices. (Gallagher 2020: 98)

My understanding of another person is action- (or interaction-) oriented, and in some cases it just is the embodied, affective, pragmatic, and intentional attitudes that pull or push me into interaction with her. When I interact with the other person I understand her in terms of my embodied preparation to respond to her, in coordination with or in opposition to her: I have immediate positive or negative sense [...]. When I see your action I see it as an affordance that motivates my own action – your action is perceived as something that I can respond to, and that is a good part of precisely how I understand your action. (Gallagher 2020: 116-117)

[...] for IT, the mind, as a second-person phenomenon, is not an internal, unobservable set of mental states. Rather the mind is “out there” in our actions and interactions, in our gestures and communications, constituted in our engagements with the world and with others. *Second-person minds* are supported by context, situations, and the social roles in which we, as agents, are engaged. In most of our everyday situations we need go no further in order to gain an understanding of the other. (Ibid.: 118).

Tramite la formulazione di questa teoria, è possibile evitare una concezione della mente esclusivamente interna all'individuo, il cui funzionamento è necessariamente da spiegare in termini rappresentazionali e i cui stati mentali devono essere inferiti dall'altro attraverso un atteggiamento basato sull'osservazione. Infatti, nella maggior parte delle interazioni quotidiane, i) la mente dell'altro è accessibile senza ricorrere a ragionamenti inferenziali o postulando meccanismi simulativi cablati nel cervello, ma direttamente tramite un processo percettivo, dal momento che gli stati mentali possono essere pensati come «dynamical embodied *processes*» (Gallagher 2020: 99), invece che come elementi statici; ii) la nostra postura di fronte all'altro in interazione non è di osservazione, quanto, invece, in seconda persona: nella maggior parte delle nostre interazioni, non siamo spettatori delle azioni dell'altro, ma contribuiamo a definirle, indirizzarle, modificarle, in quel processo di dinamico che è l'aggiustamento corporeo e interazionale<sup>99</sup>; iii) per comprendere l'altro non ci affidiamo in modo pervasivo a meccanismi inferenziali di *mindreading*, ma consideriamo le

---

<sup>98</sup> Gallagher (2020: 106) tiene a sottolineare che tale teoria non è identificabile come una forma di comportamentismo, dal momento che le espressioni corporee non sono considerate esternalizzazioni di qualcosa che è nascosto nella mente, quanto invece come elementi che permettono o alimentano il raggiungimento della formazione di una comprensione reciproca tanto sul piano sensorimotorio, quanto su quello linguistico.

<sup>99</sup> Si veda anche l'interpretazione enattivistica radicale data rispetto alla necessità dimostrata dalle neuroscienze tramite la scoperta dei neuroni specchio di portare l'analisi dell'intersoggettività dalla terza alla seconda e alla prima persona in Damiano (2010, in particolare pp. 225-228).

dinamiche interazionali come già significative, il cui senso è costruito localmente anche sulla base di incontri incarnati precedenti e sulla nostra esperienza situata nell'ambiente socio-culturale.

In sostanza, le nostre capacità percettive sono capaci di restituirci già una comprensione basata sulle azioni corporee, senza che a queste ci sia bisogno di aggiungere alcuna spiegazione inferenziale o un meccanismo di simulazione interna. Per questo, l'interpretazione enattivista dell'attivazione del pattern neuronale dei neuroni specchio non postula alcun elemento simulativo, quanto invece un evento percettivo: quando vedo gli altri compiere un'azione, i neuroni specchio che si attivano nel mio cervello non stanno simulando alcuna replica dell'azione dell'altro nella mia mente, da cui poi verrebbe inferita la sua situazione mentale, ma stanno percependo attivamente, insieme a tutto il resto del mio corpo, ciò che accade (Gallagher 2012b). Una percezione già ricca, pregnante<sup>100</sup>, quindi, e non una collezione di sensazioni provenienti dall'esterno, è ciò che permette alla teoria enattivista della cognizione sociale di porre alla propria base una radice corporea e situata della mente.

Si configura così la possibilità di rivalutare il ruolo dell'interazione e, allo stesso tempo, riuscire a far emergere l'importanza di considerare il contributo di ogni agente in essa embricato, adottando una teoria della cognizione sociale che si basa sui processi sensorimotori primari della comprensione dell'altro. Infatti, riflettere seriamente sul ruolo dell'interazione all'interno della cognizione ha almeno due conseguenze sul piano pratico, ovvero le possibilità di: i) porre come oggetto di osservazione non il modo in cui un agente si rivolge all'altro e cerca di comprenderlo in autonomia, ma il modo stesso in cui avviene la comprensione reciproca, il processo di co-aggiustamenti che rendono l'interazione sociale tale non solo in virtù delle capacità sociali degli agenti coinvolti, ma di un più profondo intreccio di percezioni e azioni, di mutue regolazioni e di creazione congiunta di significati corporei (De Jaegher 2009); ii) considerare la socialità non più come un obiettivo da raggiungere, ma come un elemento che emerge dall'interazione stessa tra due o più agenti che ne sostengono il divenire. Diventa cruciale, allora, indagare il modo in cui entrambi e ciascuno contribuiscono a sostenere o indebolire l'interazione, non solo per comprendere come la cognizione sociale si dia all'interno di questa, ma per comprendere profondamente il modo in cui la cognizione sia resa possibile da questa.

L'interazione, così, ha un ruolo contestuale, abilitante e costitutivo<sup>101</sup> (De Jaegher et al. 2010) nella realizzazione e nell'assestamento dinamico della cognizione sociale, ed è finalmente possibile

---

<sup>100</sup> Come sottolineato da De Jaegher (2009: 535), «Gallagher's claim that mirror neurons are not a mechanism of simulation but one of perception does not help because it does not dispel the idea that social cognition is something done in the individual head». Siamo d'accordo su questo e torneremo sull'argomento in seguito (§4.4.3). In questo paragrafo, tuttavia, è importante mostrare come una diversa concezione della cognizione, una che coinvolge attivamente la percezione nel processo di costruzione della comprensione, sia necessaria per oltrepassare i problemi e le questioni poste dalla ToM.

<sup>101</sup> «A contextual factor is simply something that has an effect on X, and can be determined by observing how X is changed when the factor is changed. An enabling condition not only influences the phenomenon (therefore also being

considerare lo sforzo dei singoli in un quadro d'insieme più ampio, in cui ogni elemento corporeo, pragmatico, intenzionale viene colto perché collocato nel flusso della reciprocità.

#### 3.3.4.1 Alcune conseguenze per la HRI

Ora, abbiamo visto come uno degli effetti dell'egemonia della ToM negli studi di HRI consiste nella proiezione sullo studio dell'interazione con i robot di un modello interazionale che ha la sua chiave nei meccanismi inferenziali e basato su quello che accadrebbe tra umani. L'insieme degli effetti, però, non si esaurisce qui: infatti, l'importazione di tale paradigma teorico ha comportato che, di fatto, ciò che si trova di fronte all'essere umano non avesse un gran peso in termini di considerazione all'interno dell'equazione cognitiva. In sostanza, il ragionamento innescato dall'adozione della ToM negli studi di HRI si può descrivere così: dal momento che l'essere umano opera manipolazioni di stati mentali su base inferenziale per comprendere un altro essere umano, e dal momento che il robot si presenta in interazione come se fosse un altro essere umano, ovvero *appare istaurare meccanismi sociali di interazione ed è percepito antropomorficamente*, allora l'applicazione della teoria della mente può essere utile sia a dotare il robot della capacità di interagire come un essere umano, sia a spiegare i meccanismi cognitivi che avvengono nell'essere umano. Infatti, come ricordato più volte, in questa prospettiva la mente è considerata fondamentalmente come un sistema di processazione delle informazioni che manipola rappresentazioni mentali del mondo esterno attraverso regole esplicite. Ciò è reso possibile dal fatto che le rappresentazioni mentali, ovvero quei simboli che secondo la ToM manipoliamo internamente attraverso regole per pensare, hanno un contenuto – e quindi un significato – proprio grazie alla loro capacità intrinseca di corrispondere a qualcosa nel mondo esterno. Allo stesso tempo, però, tali rappresentazioni mentali hanno una struttura funzionale, ovvero un'organizzazione interna, che ne permette l'osservazione indipendentemente dal suo rapporto simbolico con ciò che rappresentano. Di conseguenza, l'approccio cognitivista e funzionalista stabilisce che la cognizione, come fenomeno che mette insieme diversi tipi di manipolazione, possa essere studiata isolandola dal mondo in cui avviene (cfr. de Bruin & de Haan 2012).

In questo processo, fortemente antropocentrico e internalista, l'altro agente con cui ci si trova ad interagire ha effettivamente poca importanza se soddisfa alcuni requisiti minimi, dato che i meccanismi di ToM scattano di fronte a un agente che viene riconosciuto come sociale, in un modo che non permette sfumature ma solamente inferenze adeguate o meno. Quindi, partendo dalle

---

contextual), but is also necessary (either contemporaneously or historically) for X to occur. A constitutive element is part of the phenomenon (it must be present in the same time frame as the phenomenon). The set of all the constitutive elements is the phenomenon itself. The *presence* of these elements is necessary, and therefore also enabling». (De Jaegher et al. 2010: 443).



definizioni di robotica sociale e dagli effetti superficiali dell'interazione, l'obiettivo che si delinea seguendo questo approccio non è tanto capire cosa succede mentre umani e robot interagiscono, ma capire dove e come il meccanismo in oggetto si discosta o si avvicina a quella del modello cognitivo che utilizziamo per comprenderla. Il risultato è che l'interazione, a sua volta, non è un elemento necessario da considerare, dal momento che, in questo quadro, identifica solamente il risultato del modo in cui la ToM si realizza negli agenti coinvolti. Nel caso della ToM applicata all'interazione fra umani, questo non consente di osservare il modo in cui è l'interazione stessa, il suo svolgimento in vivo, in molti casi<sup>102</sup>, a determinare le azioni degli agenti coinvolti, fino a costituire la socialità della cognizione (De Jaegher et al. 2010). Ma, nell'interazione tra umani e robot, l'offuscamento è ancora più marcato, se possibile, dato che si perde la possibilità di indagare il contributo specifico del robot: infatti, guardando all'interazione solamente come il prodotto di due agenti che formulano e applicano teorie per comprendersi a vicenda, senza considerare la compartecipazione nella costruzione del processo di comprensione, si perde la specificità del robot come agente sociale, cognitivo e a suo modo *embodied* (§3.4.2), che diventa quindi solamente un'ombra sbiadita dell'essere umano, di cui vanno sottolineate le mancanze rispetto a quanto ci si aspetta, piuttosto che evidenziarne le specificità e vagliarne gli effetti.

Il cambiamento teorico che suggeriamo, allora, quello che vede il passaggio da una teoria della cognizione sociale basata sull'inferenza a una basata sull'interazione, comporta una trasformazione su tutti i fronti che interessano la HRI:

- i) abbandonare l'esclusività dell'adozione della ToM per la costruzione dell'architettura cognitiva del robot consente di procedere nella direzione aperta dalla *Nouvelle AI* e comprendere fino in fondo le implicazioni derivanti dal concentrarsi sulla dimensione sensomotoria e incarnata della cognizione sociale.<sup>103</sup> Costruire un robot che sappia coordinarsi in relazione al modo in cui noi esseri umani partecipiamo nella co-costruzione e co-gestione del significato nelle interazioni sociali necessita, dunque e prima di tutto, di una comprensione approfondita di come il robot si presenta nell'incontro con l'essere

---

<sup>102</sup> In un certo senso, sembrano sfuggire all'influenza dell'andamento dell'interazione alcune pratiche intersoggettive fortemente codificate nello svolgimento prassico. Un esempio sono tutte quelle funzioni giuridiche la cui realizzazione è vincolata da una rigida procedura, come il matrimonio: durante lo svolgimento della funzione, la libertà degli agenti è limitata dalle formulazioni che permettono alla pratica di svolgersi, pena la non realizzazione dell'atto giuridico, lasciando ben poco spazio all'improvvisazione. Tuttavia, a un livello corporeo, la coordinazione reciproca che permette il raggiungimento di un obiettivo condiviso, come in questo caso, non può dirsi immune dall'influenza dell'andamento dell'interazione: per scambiarsi le fedi, i novelli sposi dovranno comunque trovare una coordinazione corporea che permetta alla persona di infilare l'anello al dito dell'altra (e viceversa).

<sup>103</sup> Come abbiamo già sottolineato poco sopra, inoltre, riportando le critiche di Gallagher (2021) alle varie versioni della ToM, i meccanismi che caratterizzano una teoria inferenziale o simulativa della cognizione sociale sembrano comunque necessitare di una base corporea e incarnata di comprensione immediata. Gallagher elabora esplicitamente la versione dello *starting problem* in relazione alla costruzione di architetture robotiche per dotare i robot di capacità di comprensione intersoggettiva in Gallagher (2007, 2013b).

umano in quanto partecipante potenziale in un progetto di senso condiviso. Solamente dopo aver considerato gli effetti che un agente autonomo *sui generis* produce a colloquio corporeo e situato con noi, potremo delineare i presupposti del suo funzionamento interno, tale per cui il modo in cui esercita le proprie capacità si dia in stretta dipendenza alle necessità umane di co-organizzazione sociale e partecipata. Questo progetto comporta un percorso che si snoderà lungo tutto il resto della nostra tesi, che prima di tutto ha richiesto di riassetare le modalità attraverso cui vengono considerati diversi aspetti della cognizione nell'essere umano e tra esseri umani.

- ii) Infatti, abbandonare la ToM come paradigma unico e universale per descrivere ciò che accade all'agente umano nella HRI permette di dare importanza alla cognizione come processo che avviene in itinere e all'interazione dinamica, considerata come qualcosa che accade nel mondo, con dei corpi e delle materialità che si incontrano attraverso il dialogo attivo, ogni agente con le sue specifiche modalità di accoppiamento con l'ambiente. La possibilità che si configura dalla ristrutturazione dello sguardo sulla HRI è quella di ridefinire i confini dell'interazione, e con questi ricalibrare la portata degli agenti al fine di validarne il contributo singolo – ma sempre partecipato – alla formazione della reciprocità.

In sostanza, abbracciare la prospettiva enattivista alla cognizione sociale e all'interazione permette di basare lo studio della HRI non più su meccanismi inferenziali individuati per descrivere l'incontro tra umani, ma concedere all'interazione di definirsi *in vivo*, con le sue particolarità e caratteristiche realizzate proprio grazie al contributo costitutivo del corpo, della presenza e dell'esperienza di ognuno degli agenti coinvolti.

### **3.4 Interazione e nuovi agenti: per una HRI basata sulla HRI**

Ora, il cambiamento che vogliamo proporre ci permette di riconfigurare il rapporto tra interazione e singoli, in un modo che rende conto del contributo di ognuno alla costruzione del flusso interazionale e, allo stesso tempo, non considera la socialità solamente come qualcosa da raggiungere tramite l'apporto cognitivo degli agenti coinvolti e da indagare nella mente di ognuno, ma come una qualità che emerge dalla stessa pratica interazionale incarnata. Se riportiamo le considerazioni fatte nel dominio della cognizione sociale tra umani in quello della HRI, questo ci consente di sottoporre l'interazione tra umani e robot ad un'indagine differente, sotto almeno tre aspetti.

In primo luogo, l'interazione è sempre reciprocamente partecipata: il movimento di comprensione e di costruzione dell'interazione non ha mai direzionalità univoca, solamente da un agente all'altro e viceversa, ma si avviluppa nel divenire stesso dell'interazione. Nella HRI, questo implica che lo sguardo di analisi non deve concentrarsi solamente su uno dei due agenti coinvolti alla volta,

osservando come le capacità cognitive dell'umano o del robot comportino un certo tipo di cognizione sociale; attraverso lo sguardo enattivista, capiamo che l'unità di analisi deve invece essere individuata nelle modalità in cui ogni agente contribuisce all'interazione *nei confronti dell'altro*, entro un quadro complesso di scambi corporei e aggiustamenti reciproci costanti e nei rispetti della risposta alle influenze che l'interazione stessa ha sull'agente singolo.

Infatti, il secondo aspetto reso intelligibile dal cambiamento di approccio è che, oltre al fatto che l'interazione viene modellata dal contributo dei singoli, il fenomeno non si dà mai in un ambiente asettico e scevro di significato, quanto invece in un contesto ricco di esperienze passate, di affordance sociali e di predisposizioni all'azione. Una volta costituite le fondamenta dell'interazione, poi, questa ha il potere di indirizzare la cognizione di ognuno dei partecipanti. Nel dominio della HRI, questo implica che il modo in cui l'interazione si struttura risente del bagaglio di conoscenze precedenti che i due agenti riversano al suo interno, e che questa a sua volta consente di richiamare. Per il robot, questo significherà fare appello alle informazioni che il suo costruttore ha stoccato nella sua architettura cognitiva relative al funzionamento degli umani e delle loro capacità di entrare in relazione (come descritto per alcuni robot in §2.3.2); per l'umano, ciò significa percepire il robot nelle sue componenti corporee, estetiche e comportamentali, prendendo in considerazione le possibilità di incastro a partire da come il robot si presenta ed entra in interazione.

Infine, adottare un approccio enattivista alla cognizione sociale in HRI ci libera dalla rigidità teorizzata dalla ToM, secondo cui la cognizione sociale si configura solamente come attribuzione di stati mentali. Le capacità di cognizione sociale, qui, sono viste come il prodotto di numerose istanze, tutte chiamate a partecipare in diversa misura, ma che solo nel loro intreccio fanno emergere la capacità di interagire con l'altro. Nella HRI, contemplare approcci alternativi a quelli inferenziali significa aprire alle possibilità che le teorie radicali della cognizione ipotizzano riguardo l'accoppiamento tra agenti e ambienti: autonomia, *embodiment* e significato, allora sono nozioni che possono essere rese operative e applicate per chiarire quali siano le differenze tra umani e robot in interazione in termini di agentività, di intenzionalità e di *embeddedness* dell'agente nel suo ambiente.

In sostanza, possiamo raccogliere la sfida lanciata dall'enattivismo riguardo gli studi sulla cognizione sociale all'interno della HRI, e declinarla per prendere sul serio il ruolo dell'interazione nella costituzione della socialità tra umani e robot. Un problema, questo, che la HRI sembra in qualche modo già aver dichiarato la necessità di affrontare. Infatti, come sottolineato da Serge Thill e Tom Ziemke,

When interaction is considered in robotics research, however, it is often understood as two agents, one human, one artificial, each with their own cognitive apparatus, using that apparatus to engage in interaction with the other agent. This both reflects traditional views in social cognition (which are mainly interested in the individual's internal mechanisms

underlying interaction) and features the same pitfall: not explicitly recognizing that the interaction itself is fundamental, and *part of the overall cognitive process* as opposed to merely the result thereof. (Thill & Ziemke 2015: 3)

Cerchiamo, allora, di non commettere questo errore. La proposta che si vuole sviluppare consiste nel tenere a mente questi principi, che legano la costituzione dell'interazione in vivo alla natura sociale e interazionale dell'essere umano e al contributo – ancora da vagliare – che il robot apporta come variabile nel processo complesso di aggiustamenti, manipolazioni e definizioni della HRI. L'alternativa che abbiamo trovato, quella di fare affidamento alla formulazione enattivista della cognizione, permette di portare in primo piano l'analisi di ciò che accade come sforzo di entrambe le parti coinvolte, e ci consente allo stesso tempo di inquadrare la questione dell'interazione non più solamente come il prodotto dato dalla lineare somma delle parti, ma come qualcosa a cui riconoscere un'autonomia di indagine. Ciò, auspicabilmente, ci consentirà di capire quale possa essere il contributo di entrambi e di ciascuno all'interno della costruzione dell'interazione e come si possa arrivare a comprendere l'altro a partire da due punti di vista opposti e speculari. Come dichiarato da Gallagher, l'immagine che descrive meglio l'intrico di tutti i fattori coinvolti nella costruzione dell'interazione è il tango:

On the enactivist view, as embodied agents, we do not passively receive information from our environment and then create internal representations of the world in our heads; rather, we actively participate in the generation of the shared meaning we experience, the result of pragmatic and dynamical interchanges between agent and environment (Gallagher 2017; Varela et al. 1991). In the intersubjective context, interaction has an autonomy that goes beyond what the participating individuals bring to the process (De Jaegher et al. 2010). Indeed, *much like dancing the tango*, the interaction is not reducible to a set of mechanisms contained within the individual; it takes (at least) two embodied individuals who are dynamically coupled in the right way. Continuous dynamical movements between synchronized, desynchronized, and the states in-between, drive the process of social interaction (De Jaegher 2009). Attunement, loss of attunement, and re-established attunement maintain both differentiation and connection between individual agents. These processes result in the creation of meaning that goes beyond what each individual qua individual can bring to the process. In this regard, it is the interaction itself that contributes something that is not reducible to the actions of the individuals involved. (Gallagher 2020: 104, *corsivi miei*)

Comprendere, d'altronde, non è solamente «accogliere nella mente, penetrare il senso di qualche cosa», come vorrebbe la ToM, ma quando si tratta di comprendere qualcosa insieme, in una interazione che si vorrebbe sociale, comprendere è innanzitutto *comprendersi*, «avere reciproca comprensione»<sup>104</sup>, ovvero afferrarsi vicendevolmente, ma anche costruire un orizzonte di significati condivisi a partire dalle possibilità che ciascuno degli agenti ha data la sua struttura comportamentale e corporea, cioè dall'interazione.

---

<sup>104</sup> <https://www.treccani.it/vocabolario/comprendere/>

Cosa tenere in considerazione, dunque, in un approccio enattivista alla HRI? La nostra proposta, che verrà sviluppata da qui in avanti, consiste innanzitutto nell'analizzare il modo in cui il robot si presenta in interazione con l'essere umano, agli occhi di chi guarda. Questo ci consentirà di avanzare il primo passo per considerare la HRI come una disciplina che può costruire il proprio oggetto di indagine non a partire dalle concettualizzazioni fornite da altre discipline, ma individuando i propri elementi di interesse e costruendo comprensione di un fenomeno specifico laddove, invece, risultano ad oggi maggioritari gli interessi individuati dalle altre indagini. Infatti, tenere conto di come l'interazione con il robot avvenga agli occhi dell'essere umano in questa ottica compartecipata risulta necessario per basare lo studio della HRI su quello che avviene nella HRI. Sembra paradossale dover affermare quella che sembra quasi una banalità, ma abbiamo visto come il percorso aperto dalle teorie inferenziali della cognizione sociale importate nella HRI non abbia consentito, di fatto, di studiare ciò che avviene tra umani e robot senza rivolgersi a un modello di interazione che faccia riferimento a quello che succede alternativamente tra umani, tra umani e animali, o tra umani e altri tipi di macchine. La confusione e il cortocircuito generato da questo intreccio, quindi, possono essere sciolti solamente basando la HRI sulla HRI, e non su altri modelli interazionali. L'analisi della Galatea moderna deve partire da qui.

Nel capitolo precedente abbiamo visto come il robot sia considerato innanzitutto come artefatto: qualcosa che è prodotto applicando la tecnologia all'ambizione dell'umano, realizzando doppi artificiali che, da sempre, ci somigliano nelle fattezze e, dalla nascita degli automi e ancora più con la robotica, cercano di somigliarci anche nel comportamento. È questa doppia funzione del robot, allora, che adesso prenderemo in considerazione *per come si presenta in interazione*, al fine di dare conto cosa succeda di fronte al robot «in the eye of the beholder» (Brooks 1991a: 16): da una parte, la sua componente mimetica in senso estetico; dall'altra la sua componente comportamentale, nel senso sintetico. Nella visione che proponiamo, comprendere come la funzione mimetica e quella sintetica del robot si diano agli occhi dell'essere umano che si trova ad interagire con lui costituisce il primo passo per l'individuazione dei funtivi della relazione interazionale in cui l'umano e il robot, insieme, si incontrano. Comprendere i) il modo in cui il robot si presenta agli occhi dell'essere umano (§3.4.1; §3.4.2; §3.5) e ii) come l'essere umano costruisce il suo ruolo nell'interazione a partire dal riconoscimento del robot come agente sociale (§4) ci consentirà di comprendere quali sono le conseguenze di ciò che accade per la strutturazione dell'interazione nel suo complesso (§5).

Cominceremo, quindi, a vedere come il robot si presenta in quanto *artefatto mimetico* e come questo è stato concettualizzato fino ad ora all'interno degli studi di HRI, al fine di restituire una visione alternativa della mimesi del robot a partire da un account enattivista alla cognizione sociale.

### 3.4.1 *Il robot come artefatto mimetico: imitazione estetica e antropomorfismo*

Abbiamo visto che la caratteristica mostrata da automi e robot di essere artefatti che esibiscono sembianze umanoidi ha contraddistinto il loro sviluppo lungo tutta la storia dell'ambizione dell'umano (cfr. §1), compresa nella declinazione in robotica (cfr. §2). Nel tempo, e soprattutto a partire dalla nascita della robotica sociale (cfr. §2.3.2), il fatto che i robot debbano somigliare in qualche modo all'essere umano è diventata una prerogativa nella loro realizzazione, un principio fondativo (Breazeal 2004a) e una *conditio sine qua non* se lo scopo è quello di costruire artefatti che devono poter relazionarsi socialmente con gli esseri umani. Diversi studi nel campo della HRI hanno contribuito a rafforzare tale presupposto verificando la migliore efficienza dei robot che dimostrano un adeguato grado di somiglianza all'umano, soprattutto nei contesti sociali (Vélez et al. 2014). Alcuni studi evidenziano come il fatto che il robot è visto in quanto simile all'umano in termini di aspetto influenzi positivamente la sua condizione di agente sociale, incentivandone quindi accettazione e comprensione (Fink 2012; Zhao & Malle 2022). In particolare, nella ricerca sulla HRI, zone del corpo significative come il volto assumono un ruolo fondamentale nel riconoscimento del robot come potenziale agente (Lee et al. 2016; Blow et al. 2006), simmetricamente a quanto avviene nelle interazioni tra umani (Keltner et al. 2003; Ito & Urland 2005). Il robot, dunque, si configura come *artefatto mimetico* fin dal principio, innanzitutto per le sue capacità di somigliarci e di istanziare caratteristiche estetiche che ricordano quelle di un essere umano.

Una delle conseguenze principali della somiglianza del robot all'umano identificata negli studi di HRI è l'attivazione del meccanismo di antropomorfismo. L'antropomorfismo viene generalmente identificato come un fenomeno cognitivo che consiste nell'attribuzione di proprietà o caratteristiche umane ad agenti non-umani e oggetti reali o immaginari (Epley et al. 2007; Zlotowski et al. 2015a), o come il fenomeno che vede «non-human behavior as motivated by human feelings and mental states» (Airenti 2015: 121). Indagare come questa capacità cognitiva umana si innesca, è potenziata, modificata, modificabile o sfruttabile nell'interazione con i robot è uno dei principali interessi del campo disciplinare degli HRI (Damholdt et al. 2023; Roesler et al. 2021). Nella maggior parte degli studi dedicati alla tematica, il legame tra *humanlikeness* esibita dal robot a livello estetico e antropomorfismo è strettissimo e visto per lo più in maniera direttamente correlata: al crescere della somiglianza del robot all'umano, la risposta antropomorfa sarà incentivata (Hegel et al. 2008a; Rujiten et al. 2019)<sup>105</sup>.

Tuttavia, nonostante l'ampia gamma di studi e conferme riguardo al fenomeno dell'antropomorfismo, le interpretazioni riguardo l'attivazione del fenomeno, la sua struttura e le

---

<sup>105</sup> Come si vedrà in seguito (§5), alcuni studi mettono in discussione la proporzionalità diretta tra *humanlikeness* e antropomorfismo nel fenomeno dell'Uncanny Valley Effect.

cause che lo generano sono lungi dal raggiungere l'unanimità. In generale, è possibile identificare due correnti di pensiero riguardo alla natura e ai meccanismi psicologici che sorreggono l'antropomorfismo. Il confronto tra le due interpretazioni ci porterà a sviluppare una visione complessa del robot come artefatto mimetico.

#### *3.4.1.1 Non è quel che sembra: l'antropomorfismo come errore cognitivo*

Da una parte, una visione classica considera il fenomeno come «the razionalization of animal or system behavior through superimposing aspects of the human observer» (Duffy 2003: 177): in sostanza, l'antropomorfismo è qualcosa che avviene nella testa dell'osservatore umano, nel processo di identificazione, riconoscimento e comprensione di ciò che ci si para innanzi. In questa visione, l'antropomorfismo è considerato uno sbaglio innescato dalle sembianze dell'oggetto o dell'animale non umano (Guthrie & Porubanova 2020), un errore di valutazione a cui si è portati a causa della somiglianza all'umano che il robot esibisce (Turtle 2007), che può anche ostacolare un certo tipo di progresso scientifico se ciò avviene all'interno delle scienze biologiche o etologiche (Decey 2017) e che viene commesso per lo più in fasi precoci dello sviluppo (Piaget 1929) o da società primitive (Caporeal 1986; Boyer 1996). Questa visione si poggia su presupposti molto chiari, che fanno appello a una componente ontologica nei meccanismi di cognizione umana: nel caso del robot, questo artefatto *sembra* umano, ma in realtà non lo è. L'antropomorfismo è quindi il meccanismo che consente al robot di trarci in inganno e la sua *humanlikeness* è lo strumento attraverso cui la nostra percezione viene deviata e ci conduce a credere ciò che non è in realtà (Fisher 1996; Mitchell 2005; cfr. Damiano & Dumouchel 2018). Secondo un approccio classico alla HRI, il modo attraverso cui il robot ci inganna consisterebbe in una stimolazione nell'osservatore umano di meccanismi di attribuzione di stati mentali (Urquiza-Haas & Kotrschal 2015; Schilhab 2002; Spatola & Wykowska 2021), esattamente come la ToM prevede che accada nella comprensione degli altri individui nella cognizione sociale.

Questo approccio può dirsi il maggioritario all'interno della HRI, anche se non sempre ne viene esplicitata l'adesione. In questa visione, l'idea principale è quella di poter "controllare" e modulare gli istinti antropomorfizzanti dell'essere umano, elaborando innanzitutto una accurata metodologia di misurazione dei livelli di *humanlikeness* e di antropomorfismo. Come sottolineato da Voss (2021: 33): «The goal is to identify user and robot characteristics involved in the "activation" of anthropomorphic attributions». Così, il robot viene sottoposto ad un'attenta analisi e scomposto dettagliatamente nei propri connotati, in modo da poter raggiungere, attraverso un'analisi compositiva del modo in cui i vari elementi si assommano, una metrica che renda i diversi gradi e livelli di somiglianza all'umano comparabili tra di loro. Ad esempio, attributi come la taglia (Walters

et al. 2013), la colorazione e il materiale di cui il robot è composto (Wright et al. 2013), la sua forma e presenza fisica (Kidd & Breazeal 2004) vengono considerati influire sul modo in cui l'umano percepisce la *humanlikeness* e, di conseguenza, estrapolati dal contesto estetico della figura del robot per valutarne il contributo alla costruzione della qualità mimetica generale. Altri studi adottano un approccio più olistico, ma sempre basato sulla scomposizione in fattori che si compongono a formare la somiglianza all'umano, con l'intento di misurare la *humanlikeness* nel suo complesso (Bartneck et al. 2009; Ellis et al. 2005; Hinds et al. 2004; Kiesler & Goetz 2002; Kim et al. 2014; von der Pütten & Kramer 2012).

Uno degli studi più completi in merito alla *humanlikeness* esibita dai robot è quello elaborato da Phillips e colleghi (2018), i quali hanno creato un database (ABOT - Anthropomorphic roBOT<sup>106</sup>) in cui sono contenute circa 200 immagini di robot realmente realizzati che mostrino anche solo una minima somiglianza all'umano. Gli autori motivano la loro ricerca a partire dalla necessità di condurre studi quantitativi sulla *humanlikeness* dimostrata dai robot, dal momento che la mancanza di una scala oggettiva di misurazione per la somiglianza all'umano ha condotto spesso a errori di valutazione nei contesti sperimentali e a considerare robot anche molto diversi tra di loro sotto la stessa etichetta, come dimostrato per i robot descritti come umanoidi nei diversi esperimenti condotti da Stenzel et al. (2012) (a), Wiese et al. (2012) (b) e Meltzoff et al. (2010) (c).

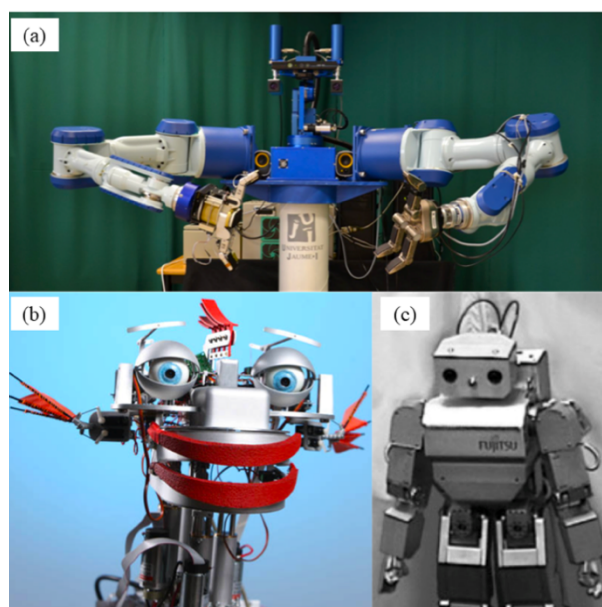


Fig. 1. da Phillips et al. (2018: 106)

A ben vedere, i robot considerati in figura dimostrano tutti e tre una qualche forma di somiglianza all'umano, ma sono comunque molto diversi tra di loro: (a) ha un torso e delle braccia, ma nessun

<sup>106</sup>Il database è aperto e consultabile al sito: <http://www.abotdatabase.info>



tratto espressivo su quello che dovrebbe essere il volto; (b) possiede dei tratti facciali molto marcati che somigliano a quelli di Kismet, ma sembra non presentare arti o corpo; (c) mostra un corpo con busto, braccia e gambe, oltre a una testa con dei tratti facciali molto stilizzati, ma comunque presenti.

Come è possibile distinguere e valutare quantitativamente questi diversi robot, che somigliano tutti all'umano ma senza dubbio ognuno in un modo differente? La metodologia dei ricercatori consiste nel classificare le diverse caratteristiche che un robot può possedere e misurarne la presenza in una scala numerica. Nello specifico, seguendo i risultati di un primo studio (Phillips et al. 2018: 107-110), vengono considerati rilevanti i) l'aspetto superficiale (*surface look*, che comprende caratteristiche come genere, pelle, ciglia); ii) l'aspetto del corpo e degli effettori (*Body-Manipulators*, che include torso, braccia, gambe); iii) le caratteristiche facciali (*facial features*, ovvero testa e occhi); iv) modalità di locomozione meccanica (*mechanical locomotion*, cioè se il movimento nello spazio è fatto su ruote, gambe, cingoli o binari). Il database ABOT viene quindi costruito valutando le sembianze del robot a partire dalle dimensioni individuate al fine verificare la rilevanza psicologica di tali dimensioni nella costituzione della impressione generale e del conseguente giudizio oggettivo riguardo la *humanlikeness* dei robot. Non sono stati ancora pubblicati studi aggiornati sul database e su come questo possa fornire una scala di valutazione oggettiva per la somiglianza all'umano nei robot<sup>107</sup>, ma già a un primo vaglio sembra che l'incrocio delle quattro dimensioni sopra individuate non conduca a una netta e uniforme classificazione dei in termini di *humanlikeness*. Infatti, come mostrato in figura 3.1, una semplice ricerca che stabilisce i parametri di ricerca per Humanlikeness (50), Body Manipulators (50), Surface look (50) e Facial features (100) restituisce un insieme di robot molto diversi tra loro e differenti in termini di somiglianza all'umano.

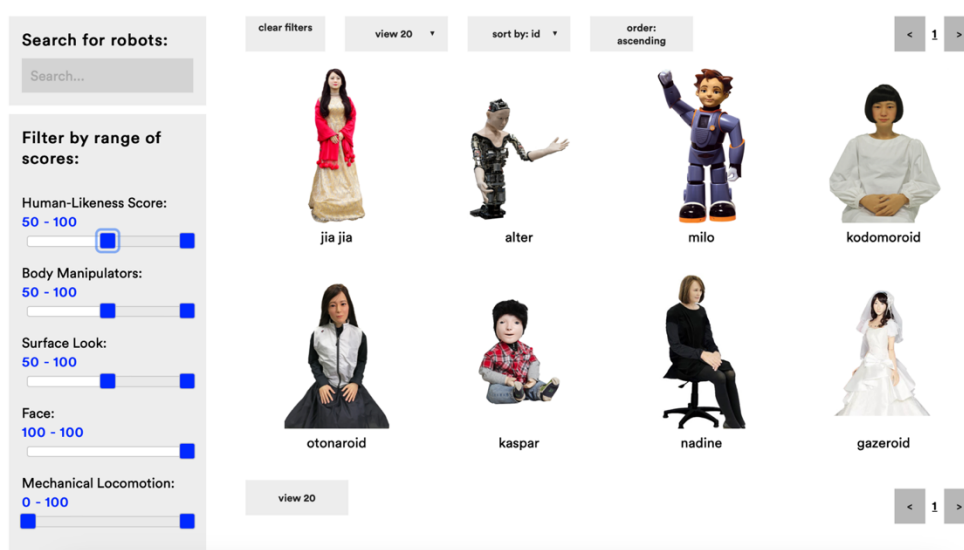


Fig. 3.1 Screenshot from <http://www.abotdatabase.info>

<sup>107</sup> Almeno alla data dell'ultima verifica, 20/01/2025.

Come è possibile che Milo (Salemi et al. 2005), un robot umanoide dai tratti quasi cartooneschi, sia categorizzato come avente gli stessi livelli di *humanlikeness* del robot Nadine (Thalmann et al. 2021), quando sembrano così diversi? Cosa accomuna Kaspar (Dautenhahn et al. 2009), con i suoi tratti così stilizzati, a Otonaroid, Gazeroid e Kodomoroid (Todo & Miisho 2014; Augey & Alcaraz 2019; Minh Trieu & Truong Thinh 2023), androidi costruiti proprio per raggiungere un aspetto al pari di esseri umani? Avere una parte dell'*hardware* scoperta, come nel caso di Alter (Maruyama et al. 2021), non lo distingue in termini di somiglianza all'umano da robot come Jia Jia (Zheng 2017), che invece indossano vestiti e non mostrano alcun meccanismo interno?

Sembra, dunque, che l'utilizzo di una dettagliata tassonomia non comporti necessariamente una maggior accuratezza nella descrizione e nella misurazione della somiglianza all'umano, né la possibilità di misurare oggettivamente il livello di *humanlikeness* esibita da ogni robot. In altre parole, esempi come questo (Lupkowski et al. 2017; Ruijten et al. 2019; von Zitzewitz et al. 2013; cfr. Mahdi et al. 2022) non riescono a catturare la complessità che risiede all'interno del concetto di somiglianza all'umano, che a quanto pare sembra irriducibile alla somma (e alla quantificazione) delle sue componenti. Anzi, a volte questo tipo di approcci arriva anche a elaborare descrizioni estreme, a partire dalla semplice presenza di tratti che di solito vengono considerati cruciali per la resa generale della somiglianza all'umano. Ad esempio, in uno studio (Kwak 2014) un'aspirapolvere automatico a cui sono stati apposti due adesivi a forma di occhi è categorizzato come "human oriented", mentre semplicemente togliendo gli adesivi la stessa aspirapolvere diventa "product oriented"; allo stesso modo, per Osawa et al. (2007), un forno elettrico con le braccia è descritto come "antropomorfo".

Lo stesso tipo di concettualizzazione sistematica, intrapresa allo scopo di fornire una spiegazione quantitativa e riuscire in qualche modo a misurarne numericamente la presenza, è stata applicata in diversi studi al fenomeno dell'antropomorfismo. Strettamente legate da un rapporto spesso causale, secondo cui un determinato grado di *humanlikeness* è concepito causare un minore o maggiore grado di antropomorfismo nell'umano, le due variabili sono confluite in alcuni test per la valutazione della risposta all'umano di fronte al robot più o meno umanoide. In particolare, uno degli studi più diffusi identifica l'antropomorfismo come una attività cognitiva dipendente da tre fattori (Epley et al. 2007), ovvero tre determinanti psicologici: l'accessibilità e l'applicabilità della conoscenza antropocentrica; la motivazione di spiegare e comprendere il comportamento dell'altro agente; il desiderio di contatto e affiliazione sociale. Anche in questo caso, si può notare come il fenomeno sia considerato come il risultato che si manifesta a partire dalla rilevazione e dal riconoscimento di alcune caratteristiche del sistema in oggetto, che si manifestano a un livello percettivo. L'adesione a tale modello concettuale comporta, da una parte, la considerazione del fenomeno come il risultato della rilevazione di alcune

caratteristiche e, dall'altra, la necessità di capire come le diverse istanze possono essere combinate per sfruttare al meglio le tendenze tipiche dell'essere umano.

Per comprendere come le istanze coinvolte portino all'antropomorfismo, una delle più diffuse modalità di misurazione in HRI è il Godspeed questionnaire<sup>108</sup> (Bartneck et al. 2009; Ho & MacDorman 2010), elaborato allo scopo di quantificare l'indice di qualità generale dell'interazione. All'interno di questo test, l'antropomorfismo ricopre un ruolo centrale, dal momento che è considerato influenzare altri aspetti presi in considerazione, quali *animacy*, *likeability*, *perceived intelligence* e *perceived safety*. Per la stima dell'antropomorfismo, il test prevede che il soggetto elabori l'esperienza di fronte al robot attribuendo una valutazione in una scala crescente che va da 1 a 5 per le seguenti coppie di qualità: Fake/Natural; Machinelike/Humanlike; Unconscious/Conscious; Artificial/Lifelike; Moving rigidly/Moving elegantly (Bartneck et al. 2009: 79). Come succede anche per altri parametri, come la *perceived intelligence* (Bacaro & Bianchini 2024), i termini coinvolti nella valutazione non sembrano catturare a pieno il significato dell'esperienza dell'antropomorfismo di fronte al robot (come nel caso della categoria semantica *moving rigidly/moving elegantly*) e tanto meno sembrano approfondire il legame tra antropomorfismo e percezione del robot come agente che somiglia all'umano (come per la categoria *fake/natural*). Inoltre, una metanalisi del test rivela che i risultati non sembrano essere omogenei se raccolti su larga scala, anche nei casi di valutazione di robot molto stilizzati come NAO (Weiss & Bartneck 2015), fallendo quindi nell'intento di costruire una scala di misurazione e valutazione valida per un'indagine uniforme e standardizzata dei concetti coinvolti. In aggiunta, il test è costruito per tenere in considerazione solamente il livello esplicito di valutazione del soggetto coinvolto nel test, tramite domande dirette e votazioni che avvengono dopo l'incontro con il robot: il ruolo del corpo, quindi, e dell'interazione non è minimamente considerato, rivelando definitivamente una concezione della percezione e dell'esperienza strettamente legata a una visione cognitivista e disincarnata della mente.

Tale impostazione comporta non solo delle conseguenze riguardo la presunta possibilità di valutare fenomeni come la somiglianza all'umano e l'antropomorfismo attraverso la sommatoria delle caratteristiche che si credano comporli, smentiti nella maggior parte dei casi da verifiche sulle metodologie di indagine, ma conduce anche a considerare fenomeni affini a quelli analizzati secondo la stessa prospettiva compositiva. Un esempio emblematico in questo senso è il modo in cui è stato inquadrato all'interno degli studi di HRI il fenomeno della deumanizzazione, preso in considerazione come corrispettivo simmetrico e speculare, uguale e contrario, dell'antropomorfismo

---

<sup>108</sup> Negli ultimi anni, i tentativi di elaborazione di test che ambiscono a diventare lo standard per la valutazione dell'antropomorfismo si sono moltiplicati. Tra questi, il RoSaS scale (Carpinella et al. 2017), HRIES (Spatola et al. 2021), Rasch model (Ruijten et al. 2019), AMS-Q (Miraglia et al. 2023). In ogni caso, il Godspeed questionnaire risulta ad oggi il più adottato (Weiss & Bartneck 2015).

(Waytz et al. 2010a; Schroeder & Epley 2016; Shin & Kim 2020). Una linea di ricerca nelle scienze cognitive, infatti, mette in relazione l'attività cognitiva di attribuire stati mentali ad oggetti inanimati e animali non umani con la negazione di tale attribuzione agli esseri umani (Haslam et al. 2007). In sostanza, tali ricerche rintracciano una base psicologica comune ai meccanismi di antropomorfismo e deumanizzazione, sfruttati però in modi diversi:

Anthropomorphism and dehumanization may then be the same underlying psychological process of mind perception, applied to different targets. If a motivation to connect to nonhuman agent increases the tendency to perceive it as humanlike, then lacking the motivation to connect with another person might lead someone to perceive a person as less humanlike (i.e., as less mindful). [...] and if bottom-up cues that suggest the presence of a humanlike mind can lead to anthropomorphism, then the absence of them when evaluating other people might lead people to see someone as relatively less humanlike. (Epley 2018: 594).

Alcuni studi sono interpretati a supporto di questa ipotesi. Ad esempio, in contesti ben identificati di in-group e out-group, alcuni partecipanti tendono a deumanizzare i membri appartenenti a un gruppo diverso in misura maggiore se si sentono inseriti all'interno del proprio (Waytz & Epley 2012); allo stesso tempo, la rimozione di tratti considerati essenziali dell'umanità, come la voce, porta a considerare l'agente meno umano (Schroeder et al. 2017). **Tuttavia, la correlazione tra antropomorfismo e deumanizzazione sembra non essere stata provata, e anzi sconfessata da altri studi (Zlotowski et al. 2017, 2014).**

Ora, abbiamo visto che, nonostante i numerosi sforzi orientati alla quantificazione, alla misurazione e alla scomposizione di fenomeni cognitivi complessi come l'antropomorfismo e la somiglianza all'umano, le evidenze mostrano una non rara difficoltà a uniformare il campo di indagine e a stabilire parametri oggettivi. Sulla scorta di quello che abbiamo visto in precedenza, possiamo tentare di comprendere perché tali studi non raggiungono la standardizzazione desiderata e la quantificazione dei fenomeni osservati.

In primo luogo, una buona parte delle difficoltà riscontrate da tali studi può essere ricondotta alla concettualizzazione teorica dei fenomeni osservati. In particolare, gli studi che abbiamo menzionato confidano tutti di riuscire a scoprire la definizione definitiva di *humanlikeness* e antropomorfismo, indagando le loro caratteristiche essenziali. Come sottolineato da Phillips e colleghi (2018: 106) nel caso della *humanlikeness*, «a precise metric is needed to compare different robots on a common scale and allow researchers to replicate findings with robots of equivalent human-likeness». In altre parole, somiglianza all'umano e antropomorfismo, nella loro concezione, possono essere definiti in modo preciso, stabile e definitivo, associando ad ognuno dei concetti delle proprietà riscontrabili nel contesto sperimentale. Tale fiducia dipende dall'adozione, non sempre esplicita, di un sostrato teorico classico computazionale della cognizione, basato sulla semantica compositiva e che vede la percezione come un processo di raccolta degli input esterni e niente di più (cfr. §2.1; §5.2.3), da cui

poi trarre le (in)adeguate conseguenze cognitive, come l'attribuzione di stati mentali a entità non umane nel caso dell'antropomorfismo. I ricercatori avviano le loro indagini con l'assunto che esista un concetto unificato di *humanlikeness* e *antropomorfismo*, caratterizzati da precisi attributi semantici che ne facilitano l'identificazione. Essi ipotizzano che, raccogliendo stimoli percettivi simili e richiamando significati astratti associati al concetto, gli individui siano in grado di discernere tra diverse entità e costruire un continuum di *humanlikeness* o di tendenza ad antropomorfizzare, stabilendo così una correlazione statistica con il loro concetto stereotipico preesistente. Le categorie di “*humanlikeness*” e “antropomorfismo” sono qui intenzionalmente costruite a partire da concetti puramente astratti, con le loro caratteristiche identificative selezionate a monte. Non c'è spazio per una nozione diversa di percezione che non sia necessariamente legata al riconoscimento di input sensoriali e di un'idea di cognizione che non sia l'elaborazione di questi stimoli per razionalizzarne l'utilizzo e computarne il significato. Lo stupore dei ricercatori quando le ipotesi non vengono supportate viene ricondotto alla poca specificazione delle caratteristiche identificative dei fenomeni (Lupkowski et al. 2017; Phillips et al. 2018), una strategia di compensazione molto simile a quella che abbiamo visto in §2.2 e che ha creato un divario tra diversi approcci nella ricerca dell'IA. Tuttavia, a partire da alcuni studi di semantica cognitiva (Rosch 1975; Rosch & Lloyd 1978), l'idea di definire i concetti a partire da una rigida categorizzazione dei tratti distintivi è stata messa in dubbio in favore di una teoria più dinamica, in cui i concetti sono riconosciuti e adottati sulla base dell'esperienza dell'individuo, del suo bagaglio culturale e del contesto di occorrenza (Violi 1997; Lakoff 1987). Il significato, quindi, non è più identificabile attraverso il reperimento dei tratti essenziali, la cui somma dà come risultato l'immediato riconoscimento del fenomeno: il significato è invece determinato dalla contingenza dell'esperienza, che seleziona elementi di un'occorrenza a partire dal contesto di realizzazione e dalle capacità interpretative del soggetto incarnato (Varela et al. 1991; Di Paolo et al. 2018; cfr. §5.2.4).

In secondo luogo, concettualizzare l'antropomorfismo come strettamente legato all'apparenza estetica del robot e al grado di somiglianza all'umano che esibisce non rende giustizia a una serie di studi che ha dimostrato la tendenza a considerare l'agente in termini antropomorfici anche se non esibisce nelle fattezze alcun tratto riconducibile all'essere umano. Infatti, a partire dallo studio di Heider e Simmel (1944), una serie di esperimenti, e numerose riproduzioni e verifiche negli anni successivi (Oatley & Yuill 1985; Berry & Springer 1993; Springer et al. 1996; Scholl & Tremoulet 2000), attesta che l'attivazione del fenomeno dell'antropomorfismo è più legato alla percezione del movimento, piuttosto che alla rilevazione di una somiglianza estetica. Nell'allestimento classico, l'esperimento prevede che ai soggetti venga mostrato un breve filmato in cui compaiono tre figure geometriche – un triangolo grande, uno piccolo e un cerchio – che si muovono in direzione diverse e

a differenti velocità. L'unica altra figura all'interno del campo visivo è un rettangolo, un lato del quale può essere chiuso o aperto. Quando i soggetti si sono trovati a dover descrivere la scena, la maggior parte di loro ha utilizzato gli stessi termini che sarebbero stati impiegati per descrivere le azioni di un essere umano, costruendo un piano narrativo in cui le azioni delle figure sono legate da una storia. Come mostrato anche da diversi studi che hanno replicato l'esperimento con bambini in età precoce (Gergely et al. 1995; Csibra 2008), gli esseri umani a partire dalla tenera età tendono a salientizzare forme che si muovono secondo un moto coerente e a interpretarlo in termini interazionali e agentivi. In altre parole, «there is no requirement of humanlikeness» (Airenti 2018: 8): senza dubbio, la presenza di tratti umanoidi incentiva l'adozione di un certo tipo di atteggiamento antropomorfizzante, ma perché istituisce un ulteriore livello di somiglianza rispetto a quello, apparentemente più fondamentale, del movimento. Trattare dunque l'antropomorfismo e la *humanlikeness* come direttamente correlate e, oltretutto, la prima come diretta e spesso unica conseguenza della qualità (o quantità) della seconda, mette in secondo piano l'importanza del movimento e dell'azione considerati essere fattori fondamentali nella comparsa del fenomeno.

Inoltre, nell'approccio classico agli studi della HRI, il fatto che la *humanlikeness* del robot causi l'antropomorfismo viene descritto come un processo di attribuzione di stati mentali a entità che in realtà non potrebbero mai raggiungere la capacità di sviluppare: il robot è solo una replica, una macchina arida di pensieri e con sentimenti artificiali, che dunque non può generare o provare pensieri e emozioni uguali o comparabili a quelli degli esseri umani (cfr. Dumouchel & Damiano 2018). Questa visione nutre l'idea che la sembianza del robot sia ingannatrice rispetto alle sue effettive capacità di rassomigliare all'umano, dal momento che queste attivano i meccanismi di attribuzione di stati mentali facendo credere che il robot possa portare avanti relazioni sociali genuine e simili in tutto e per tutto a quelle istituite con altri esseri umani. Pioniera di questa visione è Sherry Turkle, la quale sviluppa tale argomento a partire dalla concettualizzazione degli *evocative objects*, ovvero oggetti che creano «the illusion of relationship» (Turkle 2007: 514). Anche qui, vediamo come la pervasività della ToM abbia comportato un'adesione indiscussa ai modelli classici e computazionali della mente, in cui la comprensione dell'altro passa per l'inferenza dei suoi stati mentali. Tale adozione è ciò che giustifica il nesso istituito tra antropomorfismo e deumanizzazione: se il primo viene identificato come l'attribuzione di stati mentali a un'entità non umana, e la seconda con la mancanza di questa attribuzione a esseri umani, allora possono essere visti come uno il contrario dell'altro e relativi entrambi a degli errori di valutazione rispetto a un'entità che viene riconosciuta come altro rispetto a ciò che veramente è. Questa visione si poggia sull'idea che c'è una realtà oggettiva fuori dalla mente, qualcosa che funge da punto di partenza degli input e a cui gli output devono essere ricondotti affinché soddisfino delle condizioni di verità; la mente, poi, laddove ci sia,

è nascosta all'interno dell'individuo e accessibile (semmai) solamente al possessore dei pensieri, in un rapporto che può essere veritiero se i comportamenti mostrati hanno un corrispettivo interno fatto di un'altra sostanza, o menzognero se all'aspetto interno non corrisponde alcun meccanismo genuino di generazione degli effetti osservati (cfr. Damiano 2020; Damiano & Dumouchel 2018).

Infine, possiamo notare che il ruolo del corpo e dell'esperienza incarnata non è mai preso in considerazione in questi studi. L'indagine delle dimensioni di *humanlikeness* e antropomorfismo sono portate avanti attraverso la somministrazione di questionari, e soprattutto quasi sempre dopo che il robot è stato visto attraverso video o immagini. Infatti, è davvero raro che i contesti sperimentali prevedano l'interazione attiva e in tempo reale tra soggetti e robot, sia per non compromettere l'uniformità dell'esperimento e quindi la sua replicabilità, sia per via degli ingenti costi delle attrezzature necessarie (Gunes et al. 2022). Ovviamente, se vista da una prospettiva disincarnata della cognizione, questo non risulta essere un problema strutturale per indagare i fenomeni in oggetto. Tuttavia, alcuni studi sottolineano che i risultati possono essere diversi se l'esperimento viene effettuato solamente tramite videosupporti o in presenza e testimoniano dell'importanza di allestire scenari di interazione in presenza (Sciutti & Sandini 2017; Bainbridge et al. 2008; Bainbridge et al. 2011; Li 2015)

Il cambio di paradigma che abbiamo suggerito, allora, può farci comprendere come tali fenomeni non possano essere concettualizzati a partire da una loro scomposizione analitica in fattori che dovrebbero portare alla manifestazione del fenomeno, seguendo un ideale di mente basato sulla logica semantica e proposizionale e sulla percezione come collezione di input, senza inoltre tenere in considerazione il ruolo del corpo e dell'esperienza e continuando a concepire i meccanismi di comprensione dell'altro seguendo i dettami della ToM. Dobbiamo, invece, inquadrare tali fenomeni in un'ottica incarnata, in cui l'interazione allestisce possibilità di incontro e comprensione reciproca.

#### *3.4.1.2 Robot come causa ed effetto della mimesi: l'antropomorfismo come fenomeno emergente dall'interazione*

Fino a qui, abbiamo identificato la concettualizzazione classica e la inerente verifica di *humanlikeness* e antropomorfismo come indagini che si poggiano su presupposti teorici ben definiti: i) la possibilità di rendere conto dei fenomeni attraverso una loro scomposizione, che dai tratti semantici passa direttamente a configurare gli scenari sperimentali di validazioni delle ipotesi circa la quantificazione e la misurazione dei parametri; ii) lo stretto legame tra *humanlikeness* e antropomorfismo, che conduce a trascurare evidenze sperimentali alternative circa il rapporto tra antropomorfismo e movimento; iii) l'adozione di una teoria della cognizione sociale basata sulla ToM, considerata essere la teoria che spiega come avviene la considerazione dell'altro in quanto

agente umano e la sua comprensione attraverso attribuzione di stati mentali, che innesci delle conseguenze a cascata anche nella trattazione di fenomeni considerati affini come quello della deumanizzazione; iv) la validazione dei risultati attraverso questionari che non considerano il ruolo del corpo o dell'esperienza incarnata.

Possiamo, dunque, rendere conto dei fenomeni che stiamo osservando in modo diverso, se adottiamo una prospettiva enattivista come suggerito all'inizio di questo paragrafo? Come già accennato brevemente nelle pagine precedenti, il cambiamento di prospettiva su fenomeni come l'antropomorfismo ha già iniziato a comparire. Infatti, con lo sviluppo di teorie alternative della cognizione e uno studio approfondito dei processi di antropomorfizzazione del robot da parte dell'umano, sono emerse nuove interpretazioni dell'antropomorfismo, le quali si slegano dalla concezione del fenomeno cognitivo come errore di valutazione nel processo di razionalizzazione della percezione. In questa recente versione, l'antropomorfismo è considerato come una componente fondamentale della cognizione umana, una tendenza dell'organismo biologico che consente all'essere umano di vedere non solamente oggetti inanimati e agenti non umani nella loro virtuale umanità, ma soprattutto di considerarli potenziali partner in interazione.

Ciò che viene messo in luce da questo approccio è che l'antropomorfismo è una capacità cognitiva evoluta per favorire la cooperazione tra umani in ambienti sociali. Alcuni studi provenienti dall'ambito della antropologia evoluzionistica e delle scienze cognitive della religione hanno avanzato l'ipotesi che la tendenza a vedere visi e volti umani anche in forme ambigue abbia costituito un vantaggio evolutivo per le specie primitive di ominidi, il quale ha permesso alla specie di rintracciare una differenza tra alleati e nemici, di riconoscere rapidamente i predatori e di stabilire alleanze con altri gruppi di ominidi (Guthrie 1995; Bering 2005; Guthrie & Porubanova 2022). In più, studi neuroscientifici hanno mostrato come le tendenze antropomorfe degli esseri umani siano strettamente correlate all'innalzamento del livello di ossitocina negli individui (Scheele et al. 2015; Peled-Avron et al. 2016) e fortemente dipendenti dall'istinto umano alla socialità (Waytz et al. 2019)<sup>109</sup>. In sostanza, secondo questo approccio l'antropomorfismo non è più un errore o un *bias* cognitivo, ma un fenomeno che dipende strettamente dall'interesse umano nella socialità e dalle nostre capacità di interazione (Airenti 2018).

La chiave per comprendere questo fenomeno, allora, non è più quello di collegarlo a una sorta di struttura oggettiva della realtà, misurabile e quantificabile se elaborata una corretta scala di valutazione, ma individuare le circostanze e i fattori che, *all'interno dell'interazione*, favoriscono l'emersione del fenomeno negli esseri umani. In questa visione, l'antropomorfismo è il frutto della

---

<sup>109</sup> La stessa spiegazione può essere data per il fenomeno della *pareidolia* (cfr. Caruana & Seymour 2022; Akdeniz et al. 2024). Inoltre, recenti studi ipotizzano una stretta correlazione tra tendenze antropomorfe e sviluppo dell'abilità di manipolazione di utensili e della tecnica nella specie umana (Halsam 2023).



natura dialogica dell'interazione sociale, un'interazione in cui ciò che conta sono i contesti di scambio e reciprocità creati dal trovarsi in presenza di qualcosa che innesca le capacità dell'essere umano di relazionarsi proattivamente e socialmente con l'altro. In altre parole, ciò che conta nel meccanismo di antropomorfizzazione è che l'entità che si trova a essere antropomorfizzata occupa il posto di un interlocutore all'interno di una potenziale interazione umana (Airenti 2015, 2018) ed enattiva una serie di azioni che strutturano l'interazione in modo sociale (Bacaro *in press*).

Fundamentally, anthropomorphism is a way of relating with non-human entity by addressing it as it were a human partner in a communicative situation. Anthropomorphizing objects or biological entities is a means to establish a relation with them, dealing with them as interlocutors in a communicative interaction. (Airenti 2018: 8)

Seguendo questo approccio, diventa più chiaro il motivo che rende la *humanlikeness* e l'antropomorfismo due fenomeni collegati, ma non in modo causale: il fatto che l'entità mostri potenzialità interazionali simili a quelle mostrate dall'essere umano, date dal movimento, è la caratteristica sufficiente e necessaria affinché si attivi la tendenza antropomorfizzante dell'essere umano.

The fact that it is possible to attribute anthropomorphic features virtually to any object shows that resemblance is not the point. Anthropomorphism is a kind of relation that humans establish with an artifact, and in order to comprehend this phenomenon we have to focus on the relational aspect. (Airenti 2015: 117)

Non c'è bisogno di un alto grado di complessità estetica o di *humanlikeness* per innescare l'antropomorfismo, ma di un focus sulla natura relazionale della cognizione umana. Infatti, se si prende per buona la spiegazione relativa allo stretto legame tra somiglianza all'umano e antropomorfismo, non si riesce a comprendere perché gli esseri umani possono trattare la stessa entità, animale o oggetto, in modi diversi, sia in maniera "realistica" che antropomorfizzandoli (Waytz et al. 2010b; Chin et al. 2004; Cullen et al. 2014). Abbracciando una prospettiva interazionale all'antropomorfismo, d'altra parte, l'emersione della tendenza antropomorfa diventa il prodotto di una complessa serie di fattori, che non comprende solamente le qualità estetiche del robot o il *mood* dell'individuo osservato, ma si intreccia alla sua storia personale di esperienze e viene modellata dalle contingenze particolari dell'interazione in atto. Questa spiegazione fa propendere per la considerazione di certi tratti di somiglianza come fattori correlati all'emersione dell'antropomorfismo, certo altamente influenti sulla configurazione dell'interazione in ottica sociale, ma non in un rapporto di causazione diretta.

Oltre a permettere di considerare evidenze più ampie nel panorama di indagine, il disinnescare del meccanismo causale prodotto dallo stretto legame tra antropomorfismo e *humanlikeness* contribuisce anche a ripensare le modalità attraverso cui concepire la struttura del fenomeno antropomorfo in quanto fondato sulla cognizione sociale. Infatti, l'ipotesi che l'antropomorfismo sia prodotto

dall'apparenza del robot in maniera diretta non può che basarsi su meccanismi di cognizione sociale che fanno appello alla natura inferenziale della cognizione, e che l'esito 'sbagliato' a cui si giunge nel giudizio cognitivo non è altro che l'effetto della strategia di inganno percettivo impiegata dal robot: l'errore è, per così dire, giustificato dalle strategie imitative del robot.

Nella prospettiva interazionale, invece, nessun errore viene commesso, perché nessun meccanismo inferenziale è all'opera, e tanto meno la somiglianza all'umano da parte del robot viene vista come una strategia di inganno. In sostanza, l'antropomorfismo si comprende meglio se si inquadra al di fuori dell'ottica inferenziale imbastita dai meccanismi di comprensione della ToM: quando emerge il fenomeno dell'antropomorfismo, ciò che avviene è un cambiamento di ruolo dell'entità umanizzata, che passa dall'essere un oggetto tra gli altri nel mondo a qualcosa che può interagire con noi, che mostra un'autonomia e che avrà quindi la forma di un altro essere umano. Come sottolineato da Damiano e Dumouchel,

Anthropomorphic statements should not primarily be understood as descriptive statements, but as pragmatic statements in the context of interaction. As such the projection does not need to rest on the attribution of mental states to the anthropomorphized entity, nor simply any false belief. [...] Anthropomorphism is the recognition of 'intersubjectivity' in action. (Damiano & Dumouchel 2018: 5-6)

Seguendo questa prospettiva, «anthropomorphism is primarily a tool for interacting, not a description of the world» (Ivi): così, non è più una questione di comprendere come la cognizione umana si relaziona a un mondo esterno fatto di input che, quando raccolti, comportano una razionalizzazione deviante rispetto alla realtà, ma di una comprensione viva in cui la cognizione costruisce le proprie possibilità di agire in collaborazione, e non più succube, dell'ambiente che abita. Le dicotomie autentico/simulato e vero/falso su cui si poggia la prima concettualizzazione dell'antropomorfismo scompaiono per essere sostituite con una visione sistemica e dinamica dei processi cognitivi, che non implica un rapporto verocondizionale con il mondo o lo svelamento di un inganno percettivo messo in opera dall'aspetto del robot. Come esemplificato in diverse sedi da Damiano e Dumouchel (2018; Dumouchel & Damiano 2017) nel caso specifico dell'attribuzione di emozioni,

[Robots] trigger immediate emotional reactions that do not need, or rest on, the complex process of interpretation which philosophy, psychology, and classic cognitive sciences postulate as necessary for a person to access others' emotions. This affective coordination bypasses both theory of mind and folk psychology. Applied anthropomorphism does not require any false belief. (Damiano & Dumouchel 2018.: 7)

Questo quadro interpretativo sull'antropomorfismo, oltre a concedere la possibilità di indagare il fenomeno in maniera interazionale e partecipata, consente anche un arricchimento della prospettiva sulla somiglianza all'umano mostrata dal robot. Infatti, la *humanlikeness*, oltre a essere uno tra i molti

fattori, insieme al movimento, al contesto, all'esperienza dell'individuo e alle circostanze specifiche dell'interazione in corso (come la presenza o meno di obiettivi dati ai partner, di accordi specifici, di manipolazione di oggetti terzi in modalità congiunta, ecc.), che influenza l'interazione – la quale si sviluppa accentuando dinamiche in cui l'antropomorfismo è una via perseguibile e anzi preferibile – è anche concepibile come una sua conseguenza. Nel caso della HRI, il robot ci appare come artefatto mimetico non solo a priori, perché mostra delle caratteristiche che pertinentizziamo come salienti e ragguagliabili all'aspetto umano, ma anche perché nell'interazione mostra altre similarità, altri aspetti che nutrono il nostro vedere il robot *come* antropomorfo. Questa considerazione è suffragata da una serie di studi che mostrano come il fatto che il robot mostri capacità sociali e di dialogo conduce spesso all'aumento della sua *humanlikeness* percepita (De Visser et al. 2016; Cromwell et al. 2019; Fortunati et al. 2023).

Per capire meglio come avvenga il processo di aggiustamento della *humanlikeness* sulla base dell'interazione incarnata con il robot, è utile guardare alla riformulazione in ottica enattivista elaborata per il fenomeno della deumanizzazione. Negli studi classici, tale fenomeno si indentifica quando l'altro è percepito come mancante di una piena umanità, e quindi spesso considerato alla stregua di un animale o una macchina, ma a volte anche un essere senza una vita interna o un'anima (Haslam 2006), caratteristica riscontrata in casi particolari come in contesti di guerra (Bruneau & Kelly 2017) o di discriminazione razziale (Kahn et al. 2015), ma anche in momenti di quotidianità (cfr. Bastian & Haslam 2011). Secondo i sostenitori di questa visione, tale fenomeno costituirebbe una prova dell'impossibilità della comprensione diretta e primariamente corporea dell'altro in interazione, dal momento che il processo di deumanizzazione dovrebbe obbligatoriamente passare per l'elaborazione di una teoria sull'altro, informata dai pregiudizi e dalle istanze culturali a partire dal suo aspetto: in altre parole, l'evidenza dell'influenza di fattori esterni alla mente (come istanze culturali) nella comprensione dell'altro viene interpretata a favore della presenza di un meccanismo inferenziale al cuore della cognizione sociale. Come mostrato da Gallagher (2020: 150-153), tuttavia, il processo di deumanizzazione non deve necessariamente poggiarsi né sul meccanismo di (mancata o errata) attribuzione di stati mentali agli altri, né deve essere considerata *a partire* da come le fattezze dell'altro vengono percepite. Infatti, nell'insieme interazionale che si costituisce, le esperienze precedenti e le istanze culturali entrano a far parte del processo percettivo dell'altro proprio perché la cognizione sociale è prima di tutto orientata all'azione e all'interazione, e questo influisce sul modo in cui le fattezze dell'altro – e le *affordance* sociali – vengono percepite. In più, la percezione e la cognizione non devono essere visti come due meccanismi indipendenti, impenetrabili, ma come due facce della stessa medaglia in cui gioca un ruolo fondamentale il contesto culturale di appartenenza e

le esperienze vissute, in un processo che viene detto di *cultural permeation*<sup>110</sup> (Hipolito et al. 2020). Un esempio emblematico è quello in cui i volti di individui afroamericani sono considerati come più scuri di quelli di individui caucasici, anche in casi in cui il colore della pelle dei volti è stato modificato per essere lo stesso (Levin & Banaji 2006). Gallagher sottolinea che, se la percezione fosse stata informata da una teoria, come ipotizzato dagli account cognitivisti delle scienze cognitive e dalla ToM, allora, una volta che l'individuo scopre che il colore di entrambi i volti è lo stesso, l'impressione che uno sia più scuro dell'altro dovrebbe svanire, ma ciò non accade. Questo perché «we never see skin color all by itself; we see skin color along with shapes of noses and mouths – elements of a face pattern that we may associate with darker skin color because of statistical regularity» (Gallagher 2020: 153). In sostanza, non percepiamo mai una caratteristica isolata dal contesto in cui si presenta, ma percepiamo un pattern che è già significativo, il quale condiziona l'emersione delle caratteristiche che sembrano costituirlo e, al contempo, apre a significati interazionali, posture comportamentali e abiti corporei<sup>111</sup> (Varela et al. 1991).

In questo senso, allora, possiamo comparare la struttura di processi cognitivi quali l'antropomorfismo e la deumanizzazione non pensandoli come uno il contrario dell'altra, ma mettendo in luce la natura sistemica, ecologica ed enattiva di entrambi. Così come succede nel caso della deumanizzazione, nel fenomeno dell'antropomorfismo non abbiamo l'attribuzione di stati mentali a qualcosa che secondo la teoria computazionalista non potrebbe averne, ma una percezione enattiva orientata all'azione e all'interazione che permette di rendere pertinenti alcuni tratti piuttosto che altri. In sostanza, le caratteristiche estetiche del robot verranno individuate, rintracciate e percepite sotto un rispetto locale – ovvero legato alle contingenze dell'interazione e al portato esperienziale, corporeo e culturale – che elicerà l'emersione del fenomeno dell'antropomorfismo proprio in connessione alla natura intrinsecamente sociale e interazionale della cognizione umana: è per questo che si può dire che l'antropomorfismo è il riconoscimento dell'intersoggettività in azione (Damiano & Dumouchel 2018: 6). Il legame che si instaura, allora, tra *humanlikeness* e

---

<sup>110</sup> L'etichetta viene coniata per contrastare la nozione di *cognitive penetration*, basata sul concetto di cognizione come qualcosa che opera solamente a livelli alti del sistema e che può essere condizionata e influenzata (o meno) dai processi inferenziali prodotte in determinate circostanze. Nel dibattito classico, la questione concerne se i tipi di rappresentazioni mentali che stanno alla base di concetti o credenze culturalmente acquisiti possono avere un'influenza top-down che informa o condiziona il processamento di contenuti percettivi basilari (bottom-up). La risposta enattivista a questo dibattito afferma che «[...] human cognition is not only shaped but utterly permeated by the patterned practices in which it partakes and the socio-cultural contexts within which those practices grow and develop. Socio-cultural influences not only operate with respect to our explicitly formed and expressed beliefs and values, they can inform and infuse what we see and feel in quite basic ways [...]». (Hipolito et al. 2020: 164)

<sup>111</sup> «In terms of cultural permeation, not only a variety of perceptual aspects (e.g., facial patterns), as in this case, but also, more generally, cross-modal effects and action- or affordance-related changes – i.e., changes related to body schematic control – can change visual receptive fields [...]. Indeed, effects related to one's prior experience within a particular culture seem much more pervasive, and they operate in a way that modulates neuronal firing patterns, perception, bodily postures, behavioral habits, everyday practices, and intersubjective interactions (Gallagher 2017, Chapter 6)». (Gallagher 2020: 153).

antropomorfismo non è più solamente visto sotto una unidirezionalità causale che va dalla prima al secondo, ma in un ottica dinamica: un certo grado di somiglianza all'umano emerge grazie al fatto che il robot possiede alcuni tratti caratteristici dell'estetica umana e che occupa la posizione dialogica dell'altro in interazione; l'antropomorfismo è dunque *enattivato* sulla base delle esperienze precedenti di interazione sociale e porta con sé una serie di interpretazioni del robot, che rendono significativa la percezione da subito, nel suo farsi; tali interpretazioni nutrono la percezione del robot come somigliante all'umano, in un loop dinamico in cui entrambi i processi si ricalibrano vicendevolmente<sup>112</sup>.

Il ruolo del corpo e dell'esperienza è allora fondamentale, dal momento che partecipa attivamente alla costituzione del modo attraverso cui entriamo in dialogo con l'entità e costruiamo le possibilità di interazione. L'ipotesi relativa all'importanza dell'interazione in vivo si converte non più nei tentativi di parametrizzazione oggettiva e misurazione lineare dei fenomeni connessi, ma in una possibilità da cogliere per comprendere le particolarità date dalla situazione in cui l'interazione tra umani e robot sociali si svolge. Diversi studi (Damiano & Dumouchel 2018; Dumouchel & Damiano 2017; Fink 2012; Airenti 2015, 2018) testimoniano della possibilità di tenere insieme questa nuova visione dell'antropomorfismo, proponendo una nuova interpretazione del fenomeno e includendo fattori e punti di vista fino ad ora messi in secondo piano o ignorati. Ad esempio, abbracciando questo cambiamento, è possibile fornire una nuova visione sull'antropomorfismo anche nel caso della sua manifestazione linguistica, che non deve basarsi necessariamente su una correlazione che collega la produzione linguistica antropomorfa alla presenza di attribuzione di stati mentali.

If you ask a friend to borrow his jigsaw and, as he hands it to you, he adds “Be careful, it is a bit temperamental!”, how should you interpret this remark? It is unlikely that you will conclude that he sincerely believes that his jigsaw has moods, mental states and other psychological dispositions. If you do, you will have misunderstood the nature of the interaction. The point is not that this use of language is metaphorical. Rather it is that, by attributing to the other this outlandish belief, you fail to recognize what he has just done: to warn you and recommend care. [...] The anthropomorphic use of language is not metaphorical here, because there is no corresponding literal way of saying it that would accomplish what his warning and recommendation do. (Damiano & Dumouchel 2018: 5).

Applicato alla HRI, questo discorso riesce a farci considerare diversamente il modo in cui descriviamo le nostre interazioni con i robot e ci costringe a rivalutare il legame che c'è tra linguaggio antropomorfo e adozione di una postura intersoggettiva nei confronti dell'entità in oggetto. Discuteremo dettagliatamente nel prossimo capitolo (§4) la questione dell'attribuzione di stati mentali ai robot, ma per ora questo passaggio di prospettive può farci rendere conto del modo in cui riorganizzare la visione su cognizione e cognizione sociale riesca a far emergere nuove interpretazioni

---

<sup>112</sup> È necessario ricordare che il loop dinamico di ricalibramenti può sfociare, in alcuni casi, anche nell'emersione di un sentimento perturbante, come nell'Uncanny Valley Effect (§5.2).

rispetto a fenomeni chiave nel dominio disciplinare, considerati oramai quasi in maniera univoca. In questi casi, la parametrizzazione dei fattori che portano all'emersione della *humanlikeness* come caratteristica del robot e dell'antropomorfismo come re-azione alle logiche dell'interazione deve tenere conto di una molteplicità di variabili non lineari, interconnesse e dinamiche che generano il fenomeno globale dell'interazione e che retroagiscono sulla considerazione del singolo elemento aspettuale. In sostanza, si passa da una visione uni- o multidimensionale dei fenomeni a una prospettiva di interconnessione dinamica in cui il corso dell'interazione influisce sull'evoluzione e sulla manifestazione degli elementi che la compongono.

Cosa comporta, dunque, questa organizzazione per la definizione del robot in quanto artefatto mimetico? Sebbene resti a far da sfondo la caratteristica mostrata dal robot di somigliarci, questo aspetto acquista una ricchezza non solo estetica se visto in ottica enattivista, ma soprattutto interazionale: la mimesi del robot non si esaurisce sul suo volto, sulla superficie del suo corpo, ma arriva a occupare spazi di intersoggettività che possono essere colti solamente da una prospettiva incarnata e situata sul modo in cui funziona la nostra cognizione. Il robot, quindi, si configura come *artefatto mimetico* non solo perché tenta di riprodurre l'umano esteticamente, ma anche perché tramite le sue fattezze è in grado di imitare il nostro modo di relazionarci l'un l'altro e di instaurare dinamiche interazionali intersoggettive che permettono di aprire a posizioni di reciprocità tipiche dell'interazione tra umani, ricordandoci ancora una volta l'indole sociale del nostro esistere. Non solo il modo in cui vediamo il robot, ma anche e soprattutto il modo in cui percepiamo i suoi movimenti e li inquadrriamo in ottica interazionale, quindi, contribuisce alla strutturazione del suo essere oggetto *sui generis*, cosa speciale tra le altre.

C'è bisogno, ora, di un approfondimento su questo aspetto. Infatti, come abbiamo accennato in §3.4, la componente mimetica dell'artefatto robotico, almeno dalla nascita della robotica in poi, è stata accompagnata dal tentativo di replicare non solo la sembianza umana, ma anche il particolare modo che abbiamo di comportarci. Con la robotica sociale, questo ha significato indagare come riuscire a replicare il modo attraverso cui gli esseri umano si comportano con gli altri. È per questo che, necessariamente, a una disamina della componente mimetica del robot deve essere affiancata un'analisi della sua caratteristica di *artefatto sintetico*. Vedremo che sarà solo nella commistione dei caratteri mimetici e sintetici che riusciremo a rendere conto di come il robot si manifesta agli occhi di chi guarda.

### 3.4.2 *Il robot come artefatto sintetico: la questione del comportamento e il ruolo dell'embodiment*

La caratteristica che rende il robot un artefatto diverso da tutte le altre riproduzioni dell'umano è rintracciabile nel fatto che il robot si muove autonomamente, senza fili, ingranaggi e meccanismi materiali visibili, adattandosi – per quanto possibile – alle circostanze e, spesso, imparando a migliorarsi nel corso del tempo e delle interazioni. A partire da artefatti come quelli di Vaucanson e, in generale, dall'invenzione delle macchine automatiche (§1.2.1), la particolarità dell'automa è quella di muoversi, provandoci a somigliare quindi non solo nell'aspetto, ma anche nei movimenti. Quegli automi, tuttavia, non facevano emergere questioni particolarmente problematiche nei confronti del comportamento: i gesti meccanici e ripetitivi delle statue semoventi e delle bambole meccaniche erano identificabili più come una serie di azioni che come un comportamento e non mettevano in dubbio la sostanza artificiale che permetteva il loro movimento. L'aspetto più straordinario risiedeva nella loro capacità mimetica dell'umano, non nella loro capacità sintetica.

È con la rivoluzione cibernetica che la macchina diventa controparte dell'umano proprio perché reso capace di agire in modo molto simile agli esseri viventi, di esibire in qualche modo un *comportamento* che innesca descrizioni antropomorfe, che ci sembra avere obiettivi e mettere in pratica strategie più o meno efficaci per raggiungerli (Cordeschi 1998, §4). Come abbiamo visto in §1.2.3, la replicazione del comportamento generato da un essere vivente attraverso la costruzione di una macchina artificiale ha costituito la direttrice di indagine della cibernetica fin dalla sua fondazione (Wiener 1948; Cordeschi 1998, 2004). Fu proprio la cibernetica a concentrare l'attenzione non più solamente sui meccanismi interni di produzione dei processi cognitivi, ma sul legame che tali presunti meccanismi avevano con il modo in cui l'organismo/la macchina si muoveva: esempi come le tartarughe di Grey Walter (1953, §1.2.3) ben dimostrano che lo sguardo della cibernetica era rivolto principalmente a provocare un effetto nell'osservatore, lo stesso effetto che avrebbe suscitato un organismo muovendosi nell'ambiente allo stesso modo. È il metodo sintetico, marchio distintivo dell'indagine cibernetica, che mette al centro proprio la questione del comportamento: una macchina costruita artificialmente tramite i principi del metodo verrà validata solamente nel momento in cui il meccanismo implementato generi lo stesso comportamento di quello dell'organismo target. In sostanza, è l'affinamento delle tecniche di riproduzione del comportamento degli organismi nelle macchine artificiali, non più riducibile solamente a una serie di azioni, che modifica la reazione che fino a quel momento aveva connotato l'esperienza umana di fronte a un artefatto mimetico. È questo il primo aspetto che ci fa considerare il robot in quanto *artefatto sintetico*, ovvero una macchina costruita per realizzare le stesse modalità di comportamento tipiche di un organismo vivente.

Dopo che era stato messo in secondo piano dalla GOFAI (§2.1), con la svolta della *Nouvelle AI* il comportamento torna a svolgere un ruolo centrale per la costruzione di artefatti artificiali. Infatti, nell'opera di Brooks (Brooks 1991a, 1991c, 2002), lo spostamento verso una prospettiva incarnata e situata della cognizione viene motivata anche dalla necessità di riprodurre comportamenti conformi a quelli di organismi biologici anche molto semplici, come gli insetti. Grazie all'approccio *behavior-based* alla robotica (Brooks 1991b; Arkin 1998; Matarič 1998), la questione del comportamento, dall'essere relativa principalmente alle modalità di produzione dello stesso tramite implementazioni di meccanismi cognitivi come era stata nel primo periodo della robotica, si concentra di nuovo sul rapporto tra agente e osservatore, recuperando la visione cibernetica sul comportamento tramite l'evoluzione compiuta nel suo secondo periodo (§2.1.1.2). La cibernetica di secondo ordine (von Foerster 2003), infatti, a partire dallo studio dei meccanismi di regolazione e controllo dei sistemi, era arrivata a elaborare un'idea di osservatore che non è escluso dal fenomeno che osserva, ma attivamente coinvolto nel processo di definizione e comprensione del fenomeno stesso: i sistemi che osservano (von Foerster 1987) sono sistemi in cui l'osservatore riconosce che le sue descrizioni del mondo non sono oggettive e indipendenti dai fattori circostanziali, ma sono influenzate e modellate dalle sue posizioni (sia fisiche, sia epistemologiche), dagli strumenti (tecnologici o teorici) attraverso cui l'osservazione si dispiega, e soprattutto dalle sue interazioni con il sistema stesso. In sostanza, l'osservatore è parte integrante del sistema osservato, che a sua volta viene modificato da e modifica le dinamiche di osservazione e conoscenza. È grazie a questo spostamento teorico che Brooks (1991a) può affermare che l'intelligenza è negli occhi di chi guarda, ovvero in chi osserva e descrive macchine e organismi agire nei propri ambienti. Il comportamento, allora, non è più considerato solamente come il risultato di una serie di processi e meccanismi che avvengono all'interno dell'agente, nascosti e solo indirettamente inferibili, ma come un fenomeno emergente la cui adeguatezza viene rilevata in un contesto partecipato e relazionale che contribuisce alla costruzione delle impressioni dell'osservatore.

Grazie a questa concezione, individuiamo un secondo significato che è possibile raccogliere sotto la definizione di robot come *artefatto sintetico*, ovvero come qualcosa che emerge solamente dalla composizione di elementi eterogenei. Recuperando l'etimologia greca del termine, infatti, “sintesi” è quel processo di «mettere insieme»<sup>113</sup>, di combinazione che dà come risultato un processo omogeneo e intelligibile. Il robot è un artefatto sintetico non solo perché frutto dell'applicazione del metodo che porta lo stesso nome, in cui le funzioni cognitive da modellare sono *sintetizzate* in una macchina capace di riprodurle; è artefatto sintetico anche perché il comportamento che esibisce è condizionato anche dalle modalità in cui l'osservatore (e interattante) guarda il robot compiere azioni e movimenti,

---

<sup>113</sup> <https://www.treccani.it/vocabolario/sintesi/>



determinando l'adeguatezza del suo comportamento rispetto al contesto e ai propositi dell'interazione in cui si incontrano.

Come analizzare, dunque, questa doppia accezione del robot in quanto artefatto sintetico per come si dà in interazione con l'essere umano? Uno dei possibili percorsi da intraprendere consiste nel prendere in considerazione una delle caratteristiche fondamentali attraverso cui il robot è reso capace di istanziare il suo comportamento: l'*embodiment*.

Grazie alla svolta operata dalla *Nouvelle AI*, i robot costruiti secondo il modello elaborato da Brooks acquistano la capacità di poter agire nell'ambiente direttamente, e non attraverso la creazione di un surrogato interno del mondo che rappresenta l'esterno: è così che i robot diventano più efficienti, veloci e robusti nel compiere azioni che possono anche essere considerate banali, come muoversi agilmente e senza intoppi, ma che si sono rivelate problematiche da replicare per un approccio esclusivamente top-down alla cognizione e all'AI (§2.1.3). L'introduzione del principio dell'*embodiment* (Brooks 1991c) ha reso possibile che agli occhi di noi esseri umani, osservatori del sistema robotico, i robot appaiano abitare il nostro stesso mondo: possono percepire ciò che noi percepiamo, possono muoversi in accordo alla struttura del nostro ambiente, riescono a relazionarsi con noi rendendo pertinenti alcuni tratti salienti dell'interazione sociale. Lo sfruttamento dello stesso principio nella costruzione dei robot sociali, poi, ha garantito un *ground* comune su cui costruire le possibilità di interazione con gli esseri umani (§2.3.2.2).

Tuttavia, come già accennato in §2.3.3, l'applicazione a-problematica del concetto di *embodiment* agli artefatti robotici è stata fortemente criticata nel campo delle scienze cognitive che hanno riflettuto sulla robotica. Come sottolineato in diverse sedi (Sharkey & Ziemke 1998, 2000, 2001; Ziemke 2003a, 2003b, 2003c, 2004, 2007, 2016, 2022; Ziemke & Thill 2014), adottare una visione incarnata della cognizione ha reso possibile lo sviluppo di robot più efficienti, ma considerarli agenti incarnati *allo stesso modo* degli esseri umani (o di altri organismi biologici) non permette di comprendere le differenze che persistono tra questi tipi di agenti e tra le modalità in cui ognuno abita lo stesso ambiente. Rilevando tale discrepanza, Ziemke (1999, 2003a, 2003b) raccoglie una serie di interpretazioni del concetto diffuse nella letteratura delle scienze cognitive e dell'artificiale e costruisce una gerarchia di livelli di *embodiment*, ognuno relativo a una modalità specifica di prendere in considerazione l'ambiente nel processo cognitivo, e nel complesso capace di determinare la differenza tra organismi viventi e artefatti artificiali o sistemi non cognitivi:

- i) *Embodiment come accoppiamento strutturale*. Considerata la nozione più ampia di *embodiment*, identifica quei sistemi che sono strettamente legati al proprio essere situati in un ambiente. In questo senso, il sistema e l'ambiente si influenzano vicendevolmente, evolvendosi ognuno in accordo alle perturbazioni e alle contingenze dell'altro pur

mantenendo un'autonomia e un'identità. Direttamente estrapolata dalla teoria autopoietica di Maturana e Varela (1987), per quanto in quella sede la nozione veniva impiegata per descrivere il processo di co-costituzione e co-evoluzione di organismo e ambiente, Ziemke nota che l'accoppiamento strutturale non richiede la presenza di un corpo vivo e non riesce a identificare solamente i rapporti che includono sistemi cognitivi. Infatti, anche il rapporto tra i venti e le sabbie nel deserto può essere considerato sotto il rispetto dell'accoppiamento strutturale: le dune si formano, si spostano e cambiano forma in risposta alle forze del vento, mentre questo modifica la sua direzione a livello locale in risposta alle variazioni delle sabbie. Come testimoniato fin dagli insetti di Brooks, anche i robot possono essere in accoppiamento strutturale con il proprio ambiente, rispondendo alle perturbazioni e agendo attivamente nell'ambiente.

- ii) *Embodiment storico*. La nozione di accoppiamento strutturale diventa più complessa quando arricchita della caratteristica della storicità. Infatti, l'insieme delle due caratteristiche produce una descrizione dell'*embodiment* non solo derivante dallo stretto rapporto con l'ambiente in cui è l'organismo inserito, ma anche come effetto di una storia di questi accoppiamenti che si configura in un co-adattamento. In sostanza, «natural embodiment [...] reflects/embodies the history of structural coupling and mutual specifications between agent and environment in the course of which the body has been constructed» (Ziemke 1999: 187). Tuttavia, anche questa definizione non riesce a isolare la specificità dell'*embodiment* costitutivo della cognizione dei sistemi viventi: ad esempio, l'evoluzione della morfologia del letto di un fiume è senza dubbio un fenomeno che dipende dalla storia degli accoppiamenti strutturali tra il fiume e il terreno che scava per procedere. Con l'introduzione delle tecniche di *machine learning e deep learning* per le architetture robotiche (Bianchini et al. 2021; Mosavi & Varkonyi 2017), anche i robot possono essere considerati sotto il rispetto dell'*embodiment* storico dato che apprendono a modulare il proprio comportamento sulla base degli incontri con l'ambiente.
- iii) *Embodiment fisico*. Tale nozione sembra essere una delle più diffusi quando si parla di *embodiment*, considerata fondamentale almeno dalla corrente situata della robotica (Brooks 1990) e delle scienze cognitive (Pfeifer & Scheier 2001; Pfeifer & Bogard 2006). Tale nozione si distingue nettamente dal cosiddetto *funzionalismo robotico* (Harnad 1989), a cui già Searle (1980) aveva sferrato delle critiche nella *robot reply* all'esperimento della stanza cinese. Non è sufficiente costruire computer con le gambe per avere un radicamento corporeo nell'ambiente, e così raggiungere una comprensione genuina che eliminerebbe il *grounding problem*: ciò che invece è necessario per l'*embodiment* fisico è che la struttura

corporea costituisca la cognizione, ovvero che sia il modo in cui si è situati fisicamente nel mondo che dà forma ai processi cognitivi e al modo di interpretare l'ambiente tramite il comportamento. Questa nozione di *embodiment* delinea già una differenza tra organismi biologici e agenti artificiali, ma solo se si intende nella sua versione sensorimotoria (Ziemke 2003a: 1307), ovvero considerando l'apparato sensorimotorio dell'agente come costitutivo della cognizione e non solo il suo ancoramento fisico al mondo. Infatti, se si adotta il primo senso, anche i robot di Brooks (e tutti i robot costruiti per agire efficacemente nei propri ambienti senza costruzione di mondi interni) possono essere considerati come incarnati fisicamente, mentre se si adotta l'interpretazione sensomotoria e si lega la capacità agentiva all'autonomia biologica dell'organismo (Ziemke & Sharkey 2001) questi rimangono esclusi. Quest'ultima interpretazione, più restrittiva, identifica una prima differenza tra *embodiment* degli organismi viventi e quello degli agenti artificiali.

- iv) *Embodiment organismoide*. Questa nozione si concentra più che sul rapporto che c'è tra agente/organismo e ambiente, sulla *forma* che l'*embodiment* prende. Può essere considerata a sua volta un tipo particolare di *embodiment* fisico, in cui le possibilità aperte dall'accoppiamento strutturale e storico con l'ambiente sono anche condizionate dal modo in cui il corpo è fatto<sup>114</sup>. In questo senso, un corpo umano e un robot umanoide potrebbero avere le stesse possibilità di azione, dato che la forma del corpo è commisurabile.
- v) *Embodiment organismico*. Questa nozione restringe ulteriormente il campo, identificando di fatto l'*embodiment* genuino solamente con gli esseri viventi. Ziemke sottolinea che questa accezione affonda le sue radici nel lavoro di Jacob von Uexküll (Sharkey & Ziemke 2001) e trova una moderna interpretazione nella teoria dell'autopoiesi di Maturana e Varela (1980). Secondo questa visione, c'è una radicale differenza tra gli organismi viventi, *living bodies* autonomi e autopoietici, e artefatti artificiali, macchine eteronome e allopoietiche (Maturana & Varela 1982; Ziemke & Sharkey 2001). In particolare, gli organismi biologici sono macchine autopoietiche che specificano e producono continuamente la propria organizzazione<sup>115</sup> «mediante la produzione dei propri componenti, sotto condizione di continua perturbazione e compensazione di queste perturbazioni» (Maturana & Varela

---

<sup>114</sup> È in questo senso che vanno lette le ricerche sul ruolo del corpo nell'evoluzione e nello sviluppo della cognizione umana e del pensiero metaforico di Lakoff e Johnson (1980, 1999).

<sup>115</sup> Una differenza fondamentale per comprendere il contributo della teoria dell'autopoiesi è quella tra *organizzazione* e *struttura*. Per organizzazione si intende «l'insieme dei rapporti che devono esistere fra i componenti di un qualcosa perché questo possa essere considerato come appartenente a una classe particolare» (Maturana & Varela 1987: 118), mentre il termine struttura identifica «l'insieme dei componenti e dei rapporti che, concretamente, costituiscono una unità particolare nella realizzazione della sua organizzazione» (*Ibidem*). L'autopoiesi è l'organizzazione che identifica gli esseri viventi, comune a tutti gli organismi biologici, mentre la struttura varia sia tra specie (in misura massima), sia per ogni individuo della specie (in misura minima).

1972: 31)<sup>116</sup>: l'autonomia del sistema, dunque, si configura come la capacità del sistema stesso di organizzare, mantenere e riprodurre le proprie strutture interne, senza che ci sia la necessità di alcun intervento esterno e avendo come primo obiettivo «subordinare tutti i loro cambiamenti al mantenimento della propria organizzazione» (*Ivi*: 33). Al contrario, nelle macchine costruite dall'essere umano il tipo di processi espressi dall'organizzazione «non sono processi di produzione di componenti che specifichino [la macchina] in quanto unità, poiché questi componenti sono prodotti da altri processi che non partecipano alla definizione dell'organizzazione della [macchina]» (*Ivi*: 32). Questo identifica questo tipo di macchine come organizzate secondo allopoiesi, ovvero macchine che «producono attraverso il loro funzionamento qualcosa di diverso da sé stesse» (*Ivi*: 33), che non possono dirsi autonome perché non stabiliscono autonomamente i propri principi regolatori e hanno modi stabiliti dall'esterno per reagire alle perturbazioni esterne. L'*embodiment*, allora, si configura in questa accezione stretta come il modo in cui la struttura dell'organismo realizza la sua organizzazione, nel continuo aggiustamento tra precarietà del sistema e perturbazioni esterne: l'impossibilità di realizzare questo tipo di relazioni nelle macchine costituisce una delle differenze maggiori tra organismi e entità artificiali, che innesca delle conseguenze a cascata sulle possibilità cognitive del sistema osservato.

Tale classificazione, oltre a decretare l'impossibilità di equiparare la cognizione umana a quella meccanica (qualsivoglia essa sia, a meno che l'organizzazione autopoietica non sia istanziata artificialmente<sup>117</sup>), può farci rendere conto di alcune particolarità della HRI e aiutarci a delineare meglio la questione del comportamento del robot in interazione con l'umano.

In primo luogo, adottando questa gerarchia, è possibile vedere come la HRI è la sede in cui si incontrano diversi tipi di *embodiment*. Infatti, mentre l'*embodiment* proprio degli esseri umani è identificabile con il livello organismico, quello dei robot non può essere più complesso di quello organismoide. Robot umanoidi e androidi come Cog, Asimo, Sophia, Atlas, Nao o Figure, ad oggi i più complessi modelli robotici umanoidi esistenti, possono essere descritti come dinamicamente legati al proprio ambiente (*embodiment* come accoppiamento strutturale), capaci di apprendere dalle reiterazioni dell'interazione nel tempo (*embodiment* storico), aventi una struttura corporea reale

---

<sup>116</sup> «Le macchine autopoietiche sono macchine omeostatiche; la loro particolarità, comunque, non risiede in questo, bensì nella variabile fondamentale che esse mantengono costante. Una macchina autopoietica è una macchina organizzata come un sistema di processi di produzione di componenti; tali processi sono collegati tra loro in modo da produrre dei componenti che a loro volta: 1. generano i processi (relazioni) di produzione che li producono mediante le loro continue interazioni e trasformazioni e 2. costituiscono la macchina come un'unità nello spazio fisico». (Maturana & Varela 1972: 31).

<sup>117</sup> È questo uno degli obiettivi principali di una branca della biologia sintetica e dell'Artificial Life (Langton 1989; Bianchini 2023; Stano et al. 2018; Damiano & Stano 2018).

(*embodiment* fisico) simile a quella degli esseri umani (*embodiment* organismoide), ma non è presente alcuna abilità di produrre i propri componenti tramite cui si realizza la propria organizzazione e di coordinare i processi interni di sussistenza alle perturbazioni che provengono dall'esterno del sistema (*embodiment* organismico).

In secondo luogo, tramite la suddivisione dei livelli di *embodiment* si chiariscono i rapporti tra organizzazione interna e comportamento esibito dall'agente. Eteronome e allopoietiche, le macchine artificiali sono costrette dal progetto del costruttore a finite e determinate modalità di comportamento. A differenza degli organismi viventi, il cui processo perpetuo di aggiustamento che contraddistingue la loro organizzazione genera uno spazio di azioni possibili da intraprendere per garantire l'efficacia del sistema, per i robot lo spazio di manovra non è mai aperto, libero, o indagabile nel suo farsi, non è definito dalle condizioni di sopravvivenza. Le loro proprie componenti vengono assemblate secondo un progetto definito dai costruttori, che impartiscono le regole di percezione e interpretazione delle circostanze e dell'ambiente le quali, anche se modificabili nel tempo tramite processi di apprendimento, non sono guidate da una direzionalità organizzazionale strettamente legata alla sussistenza del sistema, ma da valori e parametri impartiti artificialmente; l'operatività dei robot dipende strettamente dai programmi che li compongono, non risulta dall'emersione autonoma di scopi e obiettivi, ma dall'esecuzione di programmi d'azione pre-cablati. Il comportamento del robot, allora, non può che essere eterodiretto dagli interessi formalizzati nei programmi che incorporano, e i meccanismi di apprendimento di cui sono dotati non possono che potenziare le capacità del robot seguendo binari già predisposti dai programmatori. In sintesi, il robot fa solo quello per cui è stato programmato, e riesce ad apprendere solo ciò che è stato programmato per apprendere.

È nel connubio di questi due aspetti, resi intellegibili dalla considerazione dell'*embodiment* come una gerarchia di livelli e non come un aspetto presente o assente, che è possibile delineare la questione del comportamento del robot in interazione. In HRI, l'incontro di due tipologie diverse di *embodiment* conduce a un necessario aggiustamento dell'uno sull'altro affinché l'interazione sia possibile. Nell'interazione tra esseri viventi, e specialmente nelle interazioni tra umani, problemi riguardanti la comprensione delle modalità di *embodiment* dell'altro raramente si verificano, giacché solitamente l'interazione si sostiene autonomamente sotto questo aspetto, rendendo superflua la necessità di considerare il modo in cui l'altro esperisce, percepisce e comprende l'ambiente (che comprende me, l'altro per lui) in cui l'interazione si dispiega (Gallagher 2017, 2020; Gallagher & Zahavi 2010; Zahavi 2014). Nell'interazione con i robot, invece, capire come il comportamento esibito si coordina alle modalità di *embodiment* risulta cruciale per far funzionare la comunicazione, cioè sia per interpretare adeguatamente le azioni che il robot compie, sia per costruire un piano comune di reciproci aggiustamenti in cui noi, esseri umani, possiamo compiere azioni ed essere adeguatamente

interpretati (Ziemke et al. 2015). La questione del comportamento, quindi, si configura nella dinamica di un rapporto reciproco che si istituisce in interazione e si correla direttamente alla doppia accezione di robot come artefatto sintetico: cosa può fare il robot, dato il suo tipo di *embodiment*? Cosa posso fare io, essere umano, affinché il robot possa comprendere le mie azioni?

Una versione simile della questione che abbiamo posto qui è stata recentemente elaborata nell'ambito della HRI sotto l'etichetta di *Perceptual Belief Problem* (PBP, Thellman & Ziemke 2020, 2021). Nel prossimo paragrafo, specificheremo le differenze tra il PBP e la questione del comportamento del robot come artefatto sintetico e cercheremo di evidenziare il ruolo che l'interazione incarnata può giocare per elaborare ulteriori soluzioni al problema.

#### 3.4.2.1 Il *Perceptual Belief Problem*

Il PBP, o problema della credenza percettiva, identifica «the problem of understanding what robots know (and do not know) about objects and events in the environment, based on their perception» (Thellman & Ziemke 2021: 2), anche esprimibile attraverso la domanda: «*How can people understand what robots know (and do not know) about the shared environment?*» (Ivi: 7). Gli autori caratterizzano il problema a partire da una teoria della credenza: affinché sia possibile costruire comprensione, è necessario che l'essere umano si formi le adeguate credenze in merito a cosa il robot può o non può fare, rispetto a cosa il robot può e non può percepire. In sostanza, secondo Thellman e Ziemke è importante che l'agente umano abbia un adeguato bagaglio di conoscenze e credenze che possa a sua volta guidare il processo di attribuzione di credenze, percezioni e intenzioni al robot.

I motivi per cui tale problema emerge sono individuati come direttamente conseguenti al rapporto tra ToM ed *embodiment* robotico. In particolare, nello stesso articolo gli autori sottolineano che il problema è dato: i) dalla differenza nella percezione e nella cognizione tra umani e robot, poiché non è chiaro quali tipi di cognizione possano emergere dal tipo di cognizione *diversamente embodied* del robot, che istanzia modalità percettive e cognitive diverse da quelle dell'essere umano; ii) dal fatto che il robot, grazie alla sua componente fisica, condivide lo stesso ambiente dell'interlocutore umano, ma non necessariamente le stesse capacità di percepirlo, interpretarlo e agire al suo interno; iii) dal fatto che le persone tendono a formarsi modelli inadeguati o incorretti riguardo le capacità del robot, meccanismo alimentato dalle tendenze antropomorfizzanti tipiche della cognizione umana.

Il PBP, dunque, si genera a causa del fatto che il robot è presente nel mondo, fattore che lo rende capace di agire al suo interno più di qualsiasi altro artefatto mimetico, e che si posiziona come potenziale interlocutore in una interazione sociale, cosa che secondo gli autori ha la forma dell'attribuzione di stati mentali attivati dalla sua apparenza e prodotti tramite inferenza.

Being co-present with one another, humans and robots can, in principle, see, hear, and interact with what the other is interacting with, allowing them to, among other things, anticipate each other's actions in relation to objects and events in a shared environment. However, estimating another person's knowledge based on co-presence presupposes that one is able to reliably infer the other's beliefs about objects and events in the environment. [...] In human-robot interaction such assumptions are precarious due to the fact that the mechanisms by which humans and robots perceive and understand the world differ considerably. (Thellman & Ziemke 2021: 7)

Il PBP comporta diverse difficoltà per l'interazione tra umani e robot: oltre a inficiare l'*explainability* del sistema e influenzare il legame di fiducia tra gli agenti, l'utente umano ne risente poiché non riesce a predire quali possono essere le azioni che il robot può intraprendere, minando di fatto la possibilità che l'interazione si sviluppi correttamente e seguendo un flusso spontaneo. Anche in questo caso, la previsione delle azioni del robot è descritta nei termini di formazione di una credenza e la sua validazione passa per una corrispondenza verocondizionale tra quello che l'umano crede e ciò che il robot fa.

Il ruolo della credenza e della sua attribuzione è dunque cruciale: secondo quest'approccio, è tramite la nostra credenza circa il fatto che il robot è capace di interpretare l'ambiente al nostro stesso modo che l'antropomorfismo emerge; è tramite la sua attribuzione al robot che è possibile instaurare l'interazione sociale; è dalla formazione di una credenza errata che il PBP fa emergere delle problematiche comunicative. Non si farà alcuno sforzo nel riconoscere qui lo stesso meccanismo al cuore della ToM di cui abbiamo parlato in precedenza. Anche in questo caso, infatti, il modello di cognizione sociale adottato è quello della teoria della mente<sup>118</sup>, il quale costituisce il piano teorico su cui il problema viene individuato e i termini attraverso cui viene strutturato e affrontato. I due autori fanno esplicitamente riferimento alla ToM sia nella definizione del paradigma sperimentale che adottano per verificare gli effetti del PBP, ovvero il test della falsa credenza (anche se in forma non verbale, Thellman & Ziemke 2020) sia nel definire gli stati mentali dell'altro (robot o umano) in interazione: «people's belief about other's beliefs are not directly observable. That is, they need to be inferred based on some indirect evidence» (Thellman & Ziemke 2021: 9).

Come per tutti gli altri account che abbiamo visto adottare la ToM come teoria della cognizione sociale, anche in questo caso vedere i fenomeni sotto la lente della teoria della mente implica adottare una visione internalista della cognizione. Questo è testimoniato anche dal fatto che, nonostante lo scetticismo che contraddistingue almeno uno degli autori dello studio circa la possibilità di riuscire a realizzare la complessità dell'organizzazione cognitiva umana in artefatti artificiali (Ziemke 2001; Ziemke e Thill 2014), la potenziale miglior soluzione che gli autori suggeriscono è quella di costruire

---

<sup>118</sup> Gli autori tengono a specificare che non intendono necessariamente gli stati mentali come entità proposizionali elaborate inferenzialmente (Thellman & Ziemke 2021: nota 10), ma questa precisazione fa poca differenza rispetto all'impianto generale della teoria adottata.

robot che abbiano lo stesso tipo di *embodiment*, e quindi capacità percettive e agentive, dell'essere umano: questo consentirebbe di allineare il sistema di credenze e costruire un *ground* affidabile. Notata però l'impossibilità di raggiungere questa soluzione, almeno nel breve periodo, gli autori elencano alcune alternative tramite cui oltrepassare il PBP: una soluzione endogena (interna) basata sull'esposizione prolungata dell'essere umano alla situazione problematica, nel corso della quale le persone imparano a interpretare correttamente il comportamento del robot e ad aggiustare le proprie credenze circa le loro possibilità di azione; una serie di soluzioni esogene (esterne) che si riassumono nel fornire al robot la possibilità di esplicitare esteticamente o a livello di comportamento le sue capacità e incapacità di azione, guidando così l'attribuzione di credenze percettive al robot sulla base di indizi dati dal design dello stesso (Thellman e Ziemke 2021: 10-11).

In sostanza, le soluzioni proposte riguardano i) la necessità di riequilibrare le credenze e il bagaglio di conoscenze dell'essere umano nei riguardi del robot e ii) l'allineamento delle fattezze estetiche del robot alle sue capacità percettive, o strategie compensative messe in pratica tramite avvertimenti verbali. Entrambe queste soluzioni, sebbene senza dubbio potenzialmente efficaci, comportano entrambe degli svantaggi da non sottovalutare. Per quanto riguarda i), potrebbe non essere così facile come preventivato che le persone possano avere interazioni prolungate con robot capaci di interagire socialmente. La diffusione delle tecnologie robotiche complesse, nonostante abbia subito un importante incremento negli ultimi anni, non ha ancora raggiunto un livello di pervasività tale da poter rendere disponibili robot sociali al di fuori di contesti sperimentali, di cura o educativi specialistici. D'altra parte, in merito a ii), abbassare la qualità dell'artefatto mimetico al fine di far combaciare le credenze dell'umano con le possibilità effettive del robot andrebbe a inficiare gli effetti positivi che sono stati riscontrati in letteratura circa l'efficacia di robot androidi o altamente somiglianti all'umano, ad esempio in contesti di cura (Longsdon et al. 2019; Johanson et al. 2021). In più, il fatto che il robot non mostri determinate caratteristiche specifiche in termini estetici, non implica necessariamente il fatto che l'utente umano non provi a sfruttare la particolare facoltà associata al carattere estetico mancante nel robot. Ad esempio, anche se il robot non mostra evidenti padiglioni auricolari, nella maggior parte dei casi le persone comunicano con lui tramite la voce, pensando che possa ascoltarli (Pelikan & Broth 2016; Xu et al. 2024). Per quanto riguarda la soluzione proposta di dotare il robot della capacità di comunicare direttamente in modo verbale cosa può e non può fare, ciò implicherebbe che il robot capisca cosa l'utente umano vuole da lui: nei termini proposti nello studio, ciò significa dotare il robot di una teoria della mente computazionale completa, di modo che il robot sia consapevole delle sue incapacità rispetto al contesto e alle pretese dell'umano. Come visto, tuttavia, anche questa non sembra essere una alternativa percorribile a breve termine, anche se la più battuta nell'indagine di robotica sociale e HRI (§2.3.2).



Ora, se riprendiamo i fondamenti della proposta per una cognizione sociale enattivista nella HRI, possiamo vedere a) come il problema che abbiamo indicato come la questione del comportamento sia simile ma non uguale al PBP e b) come, mettendo al centro il ruolo dell'interazione e una nozione di cognizione sociale che da questa emerge e non il contrario, possano nascere ulteriori alternative per oltrepassare il PBP.

Per quanto riguarda il punto a), il PBP e la questione del comportamento possono essere assimilati sotto il rispetto delle domande poste al sistema osservato, ovvero sulle possibilità di comprendere i meccanismi tramite cui il comportamento degli agenti coinvolti in interazione si manifesta e come riesce, se riesce, a sostenere una interazione sociale.

D'altra parte, la questione del comportamento e il PBP differiscono secondo almeno due aspetti, entrambi derivanti dall'adozione di un determinato, ma differente, concetto di cognizione. Mentre il PBP è costruito pensando alla cognizione come qualcosa di privato, i cui comportamenti altro non sono che un effetto di ciò che accade internamente al sistema, la questione del comportamento non può emergere se non strettamente correlata a una visione partecipata e dinamica dell'interazione, in cui il comportamento non è segno di qualcosa di invisibile che avviene all'interno, ma una modalità di espressione dell'agente sostenuta sia dai processi regolativi e cognitivi, sia dalla situazione interazionale nel suo complesso. Infatti, se il PBP formula il problema in termini di credenze formate e attribuite, la questione del comportamento si dispiega in relazione agli effetti sull'osservatore: fattori che non presuppongono il nascondimento dei processi cognitivi, ma che anzi si basano sulla capacità dell'essere umano di comprendere e interpretare corporalmente, nell'interazione viva, il comportamento del robot. In secondo luogo, il PBP e la questione del comportamento differiscono poiché, mentre il PBP isola una dinamica unidirezionale, che va dal comportamento del robot alla sua comprensione da parte dell'essere umano, la questione del comportamento è relativa all'andamento diadico che si instaura nel processo di comprensione, dal momento che è relativa al complesso delle dinamiche di aggiustamento realizzate nell'intreccio dei comportamenti di entrambi gli agenti. In altre parole, se il PBP vede il comportamento del robot come punto di partenza del problema, la questione del comportamento lo inserisce in un processo di aggiustamenti reciproco, la cui problematicità è individuabile solamente considerando l'interazione come un elemento costitutivo della cognizione sociale.

Queste differenze nella formulazione del problema comportano, di fatto, la possibilità di elaborare diverse soluzioni alternative, se si adotta la visione enattivista proposta (b). Come sottolineato in più articoli, non solo l'interazione ricopre un ruolo fondamentale per l'attivazione in un contesto definito di alcuni processi cognitivi, determinando in modo circostanziale lo svolgimento del compito e il modo in cui il sistema cognitivo è coinvolto, ma è un fattore costitutivo che rende possibile

l'emersione della cognizione sociale (De Jaegher et al. 2010; De Jaegher & Di Paolo 2007; Fuchs & De Jaegher 2009). Il modo in cui il PBP è formulato restringe le possibilità di soluzione o di bypassare il problema a interventi inerenti alla composizione interna della cognizione: cambiare le credenze dell'essere umano abituandolo a interazioni deludenti, modificare le modalità del robot di presentarsi e comportarsi, instradando l'interazione sui binari precostituiti dal suo tipo di *embodiment*. Aderendo a un modello che, invece, pone al centro l'interazione come elemento costitutivo della cognizione sociale, questa può a sua volta svolgere un ruolo regolatore affinché il *ground* comunicativo si costruisca proprio sulla base del ruolo duplice che i due agenti ricoprono nella costruzione dell'interazione, ogni volta sia sistema osservante sia sistema osservato. Il paradigma enattivista ci consente di considerare l'interazione come un fattore da sfruttare per dirigere gli effetti della stessa, e non solamente come un obiettivo da raggiungere. Questo ci consente di considerare come soluzione alternativa al problema non solo il miglioramento delle tecniche di replicazione del comportamento umano, ma anche di lavorare sui modelli computazionali avendo come obiettivo esclusivamente l'effetto che questo deve fare sull'osservatore. Mettere a sistema la gerarchia dei livelli di *embodiment* non con la ToM, ma con una *interaction theory* enattivista ci consente di non chiudere alle possibilità determinate dalla teoria della mente circa l'uniformità necessaria in interazione, ma contemplare la possibilità di avere interazioni su più livelli di *embodiment* che realizzano *altre* forme partecipate di socialità, quindi non necessariamente ascrivibili ai termini convenzionati della teoria della mente. Come sottolineato da Levillain e Zibetti, non c'è bisogno che il sistema robotico implementi la stessa complessità a livello cognitivo, o lo stesso *embodiment*, per generare un comportamento considerato efficace in interazione; ciò che è più importante è riuscire ad ottenere un livello di complessità di comportamento che funzioni per sostenere l'interazione sociale agli occhi di chi guarda:

A robot's behavioral complexity depends on an algorithmic sophistication that is not necessarily modeled on natural cognitive abilities. Robots do not need to possess the endogenous organization that, in human complex animals, is responsible for their behavior; what they need it to simulate the external manifestation of such an organization (Levillain & Zibetti 2017: 13).

A farci vedere meglio la questione del comportamento torna utile di nuovo l'opera di George Bernard Shaw, *Pigmalione* (1913). In questa opera, il concetto centrale è che l'adeguatezza percepita di una persona deriva dal suo comportamento esterno piuttosto che dalle competenze intrinseche che possiede. Shaw illustra come il professor Higgins riesca a convincere la società che Eliza Doolittle sia una duchessa attraverso un addestramento intensivo sul comportamento appropriato. La scommessa tra Higgins e il suo amico si basa sulla fiducia nel comportamento: se Eliza riesce a comportarsi esattamente come una nobildonna, allora la distinzione tra una vera duchessa e una finta scompare. Proprio come Eliza, un robot non necessita di possedere una comprensione intrinseca o

una coscienza umana per essere percepito come competente. Ciò che conta è la sua capacità di esibire il comportamento giusto al momento giusto.

Anche nel caso di Eliza, quindi, è negli occhi dell'osservatore che si determina l'adeguatezza del comportamento. Il bisogno di ricostruire la sua generazione ricorrendo a correlati ontologici interni alla mente è un retaggio computazionalista che è ancora vivo nella ToM, ma che con l'enattivismo possiamo lasciarci alle spalle in favore di una epistemologia relazionale e dinamica. Il comportamento è ciò che ci fa comprendere le possibilità interazionali che abbiamo di agganciarci all'altro e ci fa riconoscere un sistema come potenziale partner, il quale solitamente ha le sembianze di un essere umano nell'interazione sociale e si comporta di conseguenza, ma non esclude che altre forme di cognizione possano raggiungere risultati compatibili. Ciò che si configura, quindi, è un campo aperto di possibilità, in cui il comportamento gioca un ruolo centrale non come elemento di innesco di un meccanismo inferenziale, ma come realizzazione delle potenzialità aperte dall'interazione. È in questo modo che si manifesta la necessità di non ricondurre la HRI a modelli diversi da sé stessa, affinché sia possibile afferrare il potenziale di tale interazione nelle sue specifiche caratteristiche: la particolarità dell'interazione che stiamo trattando richiede lo sviluppo dei propri modelli di comprensione e di un impianto teorico che riesca a tenere conto delle somiglianze e delle differenze, degli aspetti in cui le possibilità di azione coincidono o meno, del modo in cui ognuno degli agenti coinvolti compone il dialogo dell'interazione e si adatta alle possibilità da questa offerte. Tramite l'adesione ai modelli della ToM, tale possibilità è preclusa fin dal principio; nel modello enattivista della cognizione sociale, invece, viene reso possibile considerare la specificità della HRI rispetto agli altri domini.

Considerare il comportamento del robot si è rivelata una necessità che ci ha condotto a spostare l'attenzione non tanto sulle caratteristiche estetiche del robot, ma sulle sue capacità di movimento e di generazione del comportamento che vengono esercitate di fronte all'essere umano. Questa linea di studi, ad oggi, risulta meno esplorata rispetto a quella che privilegia lo studio delle componenti estetiche del robot, ma abbiamo visto quanto il comportamento esibito influisca sull'essere umano, diventando determinante per afferrare il robot come compagno interazionale. Tuttavia, spostare l'attenzione sul versante del comportamento, trascurando il ruolo che la caratteristica mimetica del robot comporta nell'incontro con l'essere umano, ci condurrebbe a compiere lo stesso errore di coloro che considerano il fattore estetico del robot come più rilevante da indagare rispetto a quello comportamentale. Il riferimento all'importanza del movimento per la comprensione dell'antropomorfismo (Airenti 2015, 2018; Levillain & Zibetti 2017; Damiano e Dumouchel 2018) acquista una validità se non si prende una posizione ugualmente estrema ma contraria rispetto al legame tra capacità sintetiche e mimetiche del robot. Allo stesso modo in cui risulta incompleto

considerate in primis l'aspetto del robot come fattore scatenante dell'antropomorfismo, così la stessa inefficacia è riscontrabile nel privilegiare il ruolo del comportamento rispetto alle sembianze. Considerando il comportamento esibito come più importante da analizzare rispetto all'aspetto del robot si rischia di appiattire le differenze tra diversi agenti capaci di mostrare un comportamento simile a quello umano, ma che non esibiscono gli stessi livelli di mimesi, o *embodiment*. Per quanto riguarda i primi, alti livelli di mimesi sono fondamentali per instaurare dinamiche sociali complesse tra robot e esseri umani, come riportato in più sedi (Ishiguro 2006; MacDorman 2006a; 2006b; Balkenius & Johansson 2022); e il fatto che ci sia una importante differenza tra gli esperimenti svolti con avatar virtuali e quelli in cui il robot è in presenza e in effettiva interazione situata è stato messo in luce in diversi esperimenti, ad esempio analizzando gli effetti della presenza fisica dell'agente artificiale rispetto a quella virtuale (Bartneck 2003; Kose-Bagci et al. 2009; Leyzberg et al. 2012; Li 2015, cfr. Sciutti e Sandini 2017).

La situazione che si delinea, dunque, ci porta a dover considerare le modalità in cui le due caratteristiche del robot, mimetica e sintetica, possono essere integrate nella considerazione del singolo agente. Resta quindi da esaminare quest'ultimo aspetto, legato al rapporto tra carattere mimetico e sintetico del robot. Solo così potremmo arrivare a definire l'oggetto robot come creatura interazionale *sui generis* e affrontare la sua analisi in domini specifici.

### **3.5 Il robot come artefatto sincretico: dal modello unidirezionale all'anello interazionale**

La chiave per comprendere la particolarità del robot come artefatto consiste nel capire come le due caratteristiche che abbiamo analizzato si coordinino e si modulino in interazione con l'essere umano. A partire dall'indagine del robot come artefatto mimetico, abbiamo visto che il significato attribuibile a questa etichetta non è singolo, ma almeno duplice: infatti, il robot è artefatto mimetico non solo perché imita le nostre sembianze e ci rassomiglia esteticamente, ma perché costruisce uno spazio di intersoggettività mimando non solamente gli elementi estetici che ci contraddistinguono, replicando forme, fattezze e particolari della nostra presenza fisica, ma anche le dinamiche comportamentali che realizziamo in interazione. Questo ci ha condotto ad approfondire la caratteristica del robot come artefatto sintetico, anche qui declinato secondo una doppia accezione: sintetico perché erede diretto del metodo con lo stesso nome; sintetico perché il suo comportamento va visto come il risultato dell'accoppiamento di un sistema che osserva e di un sistema osservato, il cui grado di compatibilità è dato anche dalle modalità in cui gli agenti coinvolti abitano l'ambiente all'interno del quale sono inseriti, ovvero dal loro tipo di *embodiment*.

Come fare, allora, a tenere insieme queste due dimensioni, che ci consentono di vedere la molteplice capacità del robot di entrare in interazione con l'essere umano? Come integrare la

dimensione mimetica e quella sintetica, al fine di non perdere la specificità del robot come artefatto e basare l'indagine della HRI sulla HRI? Come riuscire a evitare di compensare l'importanza data fino ad ora al ruolo della mimesi estetica senza sbilanciare l'indagine sul versante del comportamento del robot?

A ben vedere, la soluzione può essere trovata già nel percorso che abbiamo compiuto: nello sguardo dell'osservatore, la dimensione mimetica e quella sintetica sono già integrate e si distinguono solamente nel tentativo di analizzare questo particolare artefatto nelle sue caratteristiche più peculiari; nel suo presentarsi all'osservatore, il robot è prima di tutto un *artefatto sincretico*.

Il termine 'sincretico'<sup>119</sup>, nell'accezione moderna e contemporanea, indica la composizione di elementi eterogenei che porta come risultato l'armonia degli effetti. Non troviamo termine migliore, allora, per denotare un artefatto che, nei suoi effetti, mette l'osservatore nella posizione di interagire con lui in modo che si inneschino dinamiche interazionali molto simili a quelle che avvengono con altri esseri umani. Mentre analizziamo il robot come artefatto, tenere distinti la sua apparenza e il modo in cui si muove consente di esaminare il modo in cui tali aspetti possono andare all'unisono o confliggere, dal momento che dipendono, da una parte, dal grado di integrazione tra le modalità di *embodiment* realizzate, dalla morfologia del robot, dal materiale con cui il robot è fatto, dai pattern di movimento, ecc., e, dall'altra, dalle capacità e dalle esperienze interazionali dell'essere umano, dal contesto in cui l'interazione si realizza, dalle finalità determinate localmente della stessa, e così via. Ma nell'interazione che avviene in vivo, tali caratteristiche sono meglio comprese se si considera prima di tutto che *si presentano insieme*, costruendo un effetto che alimenta l'interazione e sostiene le possibilità aperte dalla socialità della cognizione umana.

Possiamo delineare il nostro campo di indagine riguardo le influenze reciproche tra mimesi e sintesi realizzate dal robot riprendendo la gerarchia dei livelli di *embodiment* analizzata in precedenza. Il livello organismoide, livello massimo di *embodiment* possibile per una macchina al momento, è proprio quello che identifica il robot come artefatto mimetico: il robot mostra le caratteristiche corporee ed estetiche riconducibili a quelle dell'essere umano e le condizioni che rendono possibile il processo di antropomorfizzazione, sia nel processo di percezione del robot come umanoide, sia nel contesto della osservazione del suo comportamento in interazione. Tuttavia, il fatto che *la forma dell'embodiment* sia condivisa non comporta l'annullamento delle differenze in termini di cognizione, dal momento che a simili forme di *embodiment organismoide* non necessariamente corrispondono le stesse possibilità di impiego del corpo nel dispiegarsi del comportamento. Il fatto che il robot istanzi un tipo organismoide di *embodiment* costruisce la possibilità che lo stesso sia considerato come potenziale interattante. Il fatto però che il robot si fermi a questo livello, senza poter

---

<sup>119</sup> <https://www.treccani.it/vocabolario/sincretismo/#>

di fatto raggiungere il livello organismico che contraddistingue ambedue gli agenti nell'interazione tra umani, rende necessario connotare l'interazione come tendente all'umano, ovvero antropomorfica, e non antropica.

È in questo elemento di scarto, in questo intervallo tra diverse forme di *embodiment* che il robot si caratterizza come artefatto sincretico i cui effetti si realizzano in un'interazione che si verifica a cavallo tra *embodiment* orgasmoide e organismico. Questo non preclude la possibilità di avere come effetto la realizzazione di una qualche forma di socialità. Come abbiamo sottolineato in precedenza (§2.4), l'enattivismo apre alle possibilità che restano limitate se si adotta l'idea computazionalista e funzionalista di cognizione: a seconda dei gradi di *embodiment* che contraddistinguono gli agenti coinvolti in interazione, si può arrivare ad ottenere anche una forma di *embodiment* sociale (Ziemke 2003a), ovvero quel tipo di relazione con il proprio ambiente che coinvolge il corpo nelle interazioni sociali e in cui la situazione fisica dell'agente influenza ed è influenzata dalle interazioni vive e partecipate con gli altri agenti (Vernon 2014: 126-127). La natura sincretica dell'artefatto robotico, dunque, si esprime anche nelle possibilità aperte dall'enattivismo di poter concepire interazioni sociali in cui diverse modalità di istanziare il comportamento si trovano e si incontrano producendo effetti sulle possibilità di aggiustamento e reciprocità. Possiamo esprimere questo concetto tramite le parole di Luisa Damiano, che connota la possibilità aperta dall'enattivismo come una potenzialità di co-evoluzione nell'interazione tra umani e robot.

L'enazione, negando ai robot l'autonomia di auto-produzione tipica dei sistemi biologici, rifiuta loro i peculiari tratti cognitivi e sociali in cui essa si esprime, nello specifico, al livello antropologico. Allo stesso tempo, riconosce gli agenti robotici detti sociali come sistemi che, in base a prestazioni cognitive e sociali di carattere computazionale, in certe condizioni possono coinvolgere gli umani in specifici processi di "accoppiamento comportamentale". È la possibilità teorica di co-evoluzioni uomo-robot in cui agenti umani e agenti robotici, mediante scambi di segnali sociali, rendono interdipendenti alcuni aspetti delle loro rispettive dinamiche e disposizioni all'azione, mantenendo o potenziando la propria efficacia operativa nelle ecologie di riferimento. [...] L'idea è quella di forme estremamente limitate e specifiche, ma effettive, di accoppiamento comportamentale, la cui influenza sulla nostra condotta richiede studi specifici. (Damiano 2020: 35)

Un accoppiamento comportamentale, quindi, è quello che contraddistingue il robot come artefatto sincretico e che necessita di studi specifici per comprendere al meglio come realizzarne di efficienti nell'interazione con l'umano. Vedere il robot sotto la lente del suo sincretismo è quello che ci permette di spostare il focus dai modelli di cognizione individuali, in cui si cerca di comprendere la cognizione sociale come un caso particolare di ciò che avviene in isolamento dai contesti interazionali, a un modello costitutivamente partecipato, un *anello interazionale* in cui gli effetti non possono essere compresi se non passando per la complessità delle relazioni potenziali e attuali comprese nell'incontro con gli altri.

Ricapitolando, il sincretismo dell'artefatto robotico può essere riassunto così:

- i) Le sue caratteristiche di artefatto mimetico e sintetico non possono essere esaminate come fenomeni incapsulati e nettamente distinti, ma devono esseri analizzati nell'ottica compartecipata dell'interazione reciproca che produce l'armonia degli effetti agli occhi dell'osservatore, per cui il robot è visto come interlocutore sociale.
- ii) L'artefatto sincretico non trova il suo potenziale sociale nella realizzazione di un meccanismo computazionale efficace perché condiviso con gli esseri umani, ma perché si colloca come interattante in cui l'intreccio di capacità mimetiche e sintetiche enattiva gradienti di socialità e in cui l'interazione gioca un ruolo cruciale per definire i parametri dell'aggiustamento reciproco e coordinato.

Si apre allora una sfida relativa a come trattare il robot come artefatto sincretico non trascurando alcuno degli aspetti che questo importa nell'interazione. Comprendere il modo in cui le due caratteristiche di artefatto mimetico e sintetico si intreccino a comporre il sincretismo del robot come interattante è risultato cruciale per afferrare la specificità della interazione della HRI, ma si configura come ancora più rilevante se consideriamo l'operazione di costruire artefatti robotici capaci di agganciarsi alle possibilità aperte dall'accoppiamento comportamentale su cui l'interazione tra umani e robot si modella. Comprendere come il robot in quanto artefatto sincretico si manifesta in interazione agli occhi dell'essere umano comporterebbe la possibilità di basare la costruzione di robot sociali sulle dinamiche di aggiustamento effettive, concrete, incarnate, sociali, partecipate, reciproche e situate in cui si svolge l'interazione sociale, e non su modelli plausibili di interazionalità decisi a monte.

La *Galatea moderna* che ci siamo proposti di analizzare all'inizio di questo capitolo incarna una rivoluzione nel modo in cui concepiamo e interagiamo con i robot. Seguendo una visione enattivista, che elabora una versione alternativa alla ToM come spiegazione della cognizione sociale e si concentra invece sulla partecipazione incarnata degli agenti, emerge una nuova visione delle interazioni umano-robot. Questo paradigma ci invita a vedere i robot non come semplici strumenti programmati per eseguire compiti, e non solo come supporti su cui provare a realizzare le stesse modalità cognitive che si sviluppano in interazione con gli altri; al contrario, una concezione radicale della cognizione sociale ci consente di vedere i robot come partner attivi e dinamici con cui co-creiamo significato e comprensione. La partecipazione incarnata suggerisce che la cognizione non è un processo isolato all'interno di una mente, ma una pratica condivisa ed emergente dall'interazione continua tra esseri umani e macchine. Così, la *Galatea moderna* diventa un simbolo della possibilità

di una sinergia autentica e significativa, un invito a esplorare nuovi orizzonti nella nostra co-evoluzione con la tecnologia (cfr. §5.4).

Allora, se è vero che nell'ottica enattivista «interaction with complex, but not fully autonomous systems (such as [...] social robots) can provoke an experience of engagement» (De Jaegher et al. 2010: 442), ciò che dobbiamo capire è *come* avvenga questa partecipazione, come si dia in sostanza la collaborazione nella co-costruzione dell'incontro interazionale tra agenti con modalità differenti di realizzazione della cognizione sociale. Il fenomeno chiave su cui testare le nostre modalità di comprensione del robot come artefatto sincretico è uno dei più discussi attualmente all'interno del campo della HRI, ovvero l'attribuzione di intenzionalità ad agenti artificiali. Nel prossimo capitolo, affronteremo l'analisi di questo fenomeno per i) comprendere come l'artefatto sincretico si dà in interazione, agli occhi di chi guarda, ii) per costruire un modello di analisi e comprensione della HRI basata su un approccio enattivista alla cognizione sociale e iii) per informare la costruzione robot sociali che riescano a sostenere l'accoppiamento comportamentale in modo efficace.



## Capitolo IV – Umani, troppo umani.

### Il ruolo dell'intenzionalità nell'interazione tra umani e robot

Nous voudrions précisément montrer que le robot n'existe pas, qu'il n'est pas une machine, pas plus qu'une statue n'est un être vivant, mais seulement un produit de l'imagination et de la fabrication fictive, de l'art d'illusion. Pourtant, la notion de la machine qui existe dans la culture actuelle incorpore dans une assez large mesure cette représentation mythique du robot. Un homme cultivé ne se permettrait pas de parler des objets ou des personnages peints sur une toile comme de véritables réalités, ayant une intériorité, une volonté bonne ou mauvaise. Ce même homme parle pourtant des machines qui menacent l'homme comme s'il attribuait à ces objets une âme et une existence séparée, autonome, qui leur confère l'usage de sentiments et d'intentions envers l'homme.<sup>120</sup>  
(Simondon 1958: 11)

In questo passaggio di *Du mode d'existence des objets techniques*, Simondon riassume in poche righe la questione che andremo ad affrontare in questo capitolo. Infatti, riallacciando la figura del robot alle sue radici mitiche e finzionali, il filosofo francese descrive la figura della macchina semovente come qualcosa che *non esiste* per sé stessa, in quanto tale, per lo meno non più di quanto una statua esista come essere vivente. Ciò che Simondon mette in luce è che, nonostante la natura di questo tipo di macchina non sia altro che quella di un artefatto dalle capacità *mimetiche* (e noi possiamo aggiungere *sintetiche*), l'essere umano è portato a vedere il robot come guidato e mosso da motivazioni, a considerarlo come un agente al suo pari, a ipotizzare che ci sia qualcosa in più dei soli meccanismi e ingranaggi (o algoritmi e intelligenze artificiali) ad animarlo. È la forza dell'immaginazione, ci dice Simondon, a spingerci a vedere nel robot non solo un artefatto tecnico, ma un'entità capace di volontà e azioni guidate da scopi.

Nonostante sia solo una macchina, e quindi incapace di ragionare, percepire, agire volontariamente come un essere umano, c'è qualcosa che ci spinge a considerare il robot diversamente, a usare un linguaggio simile a quello che adottiamo per descrivere altri essere umani o entità vive, con qualcosa

---

<sup>120</sup> Traduzione mia: «Vorremmo dimostrare con precisione che il robot non esiste, che non è una macchina, così come una statua non è un essere vivente, ma solo un prodotto dell'immaginazione e della fabbricazione fittizia, dell'arte dell'illusione. Tuttavia, la nozione di macchina presente nella cultura attuale incorpora in larga misura questa rappresentazione mitica del robot. Una persona colta non si permetterebbe di parlare degli oggetti o dei personaggi dipinti su una tela come se fossero realtà vere, dotate di interiorità, di una volontà buona o cattiva. Eppure, quella stessa persona parla delle macchine che minacciano l'uomo come se attribuisse a questi oggetti un'anima e un'esistenza separata, autonoma, che conferisce loro l'uso di sentimenti e intenzioni nei confronti dell'uomo».

– come un’anima – che lo muove dal di dentro, anche quando sappiamo che il suo comportamento è diretto da un insieme di meccanismi o algoritmi. Di fatto, molti studi che indagano diversi fenomeni, come antropomorfismo e percezione della macchina come agente intenzionale, testimoniano che il robot viene descritto come dotato di scopi, pensieri, emozioni e volontà, descrizioni che denotano una postura intenzionale nei loro confronti. Ma non solo: un diverso insieme di studi testimonia che il robot può venir trattato in maniera simile a quanto succede con altre entità autonome, con atteggiamenti che ricordano quelli esibiti di fronte a esseri viventi o umani. Come è possibile, allora, che i robot riescano a eludere il limite strutturale posto dalla loro organizzazione artificiale fino a suscitare nell’essere umano la tendenza a vederli, a trattarli, descriverli come veri e propri agenti? Come avviene questo processo?

Il dilemma posto da Simondon ha acquisito rilevanza nel dibattito sull’interazione tra umani e robot, assumendo la forma del *problema dell’attribuzione di intenzionalità*, oggi considerato uno dei nodi teorici principali nel campo di studi (Ziemke 2023). La concentrazione dell’interesse sulle dinamiche relative al fenomeno della percezione dell’intenzionalità del robot ha aggiunto nuove variabili alla condizione dell’umano in questa esperienza: oltre a dare conto sperimentalmente delle intuizioni di Simondon, si è anche scoperto che non sempre l’essere umano adotta un linguaggio antropomorfo di fronte al robot, soprattutto quando deve rispondere a domande dirette sulla presenza di un’anima, di una mente o di una volontà: accade anche che, a volte, a fronte di un linguaggio che presenta termini che potrebbero far ipotizzare la considerazione del robot come intenzionale, il comportamento di fronte a loro non è identificabile con questa posizione. Il nodo, allora, si fa ancora più intricato.

In questo capitolo cercheremo di mappare il problema attraverso l’analisi del concetto di intenzionalità nel panorama storico e contemporaneo delle scienze cognitive (§4.1), considerando le modalità in cui tale concetto viene impiegato nello studio dell’attribuzione di intenzionalità ai robot sociali e l’impianto sperimentale allestito per verificarne la presenza (§4.2). A seguito di alcune riflessioni relative al modo in cui le nozioni di intenzionalità e della sua attribuzione vengono declinate negli studi di HRI, verrà mostrata la necessità di spostare il focus da un meccanismo di attribuzione di intenzionalità ai robot sociali, basato sull’adesione alla ToM come approccio alla cognizione sociale, a un processo di rilevazione, identificazione e riconoscimento (*detection*) dell’intenzionalità nel comportamento del robot, basato su un approccio enattivista alla cognizione sociale (§4.3). In questo modo, sarà possibile offrire un’interpretazione alternativa degli esperimenti sull’attribuzione di intenzionalità ai robot, evitando l’uso improprio di concetti ibridi e aprendo nuovi orizzonti per comprendere il fenomeno del rilevamento dell’intenzionalità in termini corporei e situati. Questa operazione ci condurrà a ripensare anche i termini con cui definiamo il problema che

stiamo analizzando. Il movimento che va da un approccio mentalistico e inferenziale a uno incarnato ed enattivo comporta anche una riconfigurazione del processo che coinvolge l'intenzionalità: dalla *attribuzione* di intenzionalità al suo *riconoscimento*. A questa riconfigurazione farà seguito una riflessione riguardo al ruolo che una visione enattivista dell'intenzionalità può avere negli studi sull'interazione tra umani e robot: sarà necessario esplicitare alcuni assunti interni a tale concetto affinché una nozione radicalmente enattivista di intenzionalità possa emergere, supportata da alcuni studi provenienti dalla ricerca in HRI (§4.4).

In sintesi, l'obiettivo di questo capitolo può dirsi in linea con quanto formulato da Simondon: si vuole dimostrare che il robot non esiste *per sé stesso, in quanto tale*; è solo se visto nell'ottica compartecipata dell'interazione che lega essere umano e robot che quest'ultimo può dimostrare le capacità che non possiede a livello di organizzazione e acquistare le qualità che sembrano trascendere la sua mera struttura tecnica.

Il primo passo di questo processo di riformulazione consiste, dunque, nel chiarire i termini e i concetti fondamentali di questo dilemma: cosa intendiamo quando parliamo di intenzionalità?

#### **4.1 Intenzionalità: il marchio del mentale**

L'intenzionalità viene generalmente definita come la caratteristica della mente di essere diretta verso qualcosa, di essere circa (*about*) un oggetto o uno stato di cose, come inteso nella sua radice etimologica latina (*tendere*, Jacob 2003). La centralità di questo concetto per comprendere la natura della mente viene sancita dallo psicologo e filosofo tedesco Franz Brentano, che recupera una riflessione classica sviluppatasi a partire dagli scritti medievali di Tommaso d'Aquino. Si può affermare senza timore di smentita che qualsiasi riferimento moderno e contemporaneo all'intenzionalità all'interno delle scienze cognitive parte dalla definizione che Brentano fornisce del termine:

Ogni fenomeno psichico è caratterizzato da ciò che gli scolastici medioevali chiamarono l'in-esistenza *intenzionale* (ovvero *mentale*) di un oggetto, e che noi, anche se con espressioni non del tutto prive di ambiguità, vorremmo definire il riferimento a un contenuto, la direzione verso un oggetto (che non va inteso come una realtà), ovvero l'oggettività immanente. Ogni fenomeno psichico contiene in sé qualcosa come oggetto, anche se non ciascuno allo stesso modo. Nella presentazione qualcosa è presentato, nel giudizio qualcosa viene o accettato o rifiutato, nell'amore qualcosa viene amato, nell'odio odiato, nel desiderio desiderato, ecc.

Questa in-esistenza intenzionale è esclusiva caratteristica dei fenomeni mentali. Nessun fenomeno fisico esibisce qualcosa di simile. Possiamo dunque definire i fenomeni mentali come quei fenomeni che contengono un oggetto in modo intenzionale. (Brentano 1874: 164-166)

Non ci sembra fuori luogo richiamare l'immagine dei proverbiali fiumi di inchiostro versati per scrivere commenti a tale definizione ed elaborarne interpretazioni. Infatti, in questo densissimo passo è possibile scorgere almeno due delle questioni che hanno animato il dibattito riguardo la natura dell'intenzionalità, il suo ruolo nella comprensione della mente e sulle possibili modalità di descrivere il processo intenzionale, temi che risultano ancora oggi centrali nelle scienze cognitive e nelle sue applicazioni all'interazione con gli artefatti artificiali.

In primo luogo, l'intenzionalità viene descritta come una proprietà della mente non riducibile o spiegabile in termini di fenomeni naturali, dal momento che il dominio psicologico e quello fisico sono ontologicamente diversi, fattore che istituisce una incomunicabilità diretta fra i due. Tale affermazione, che non suonerà completamente inedita (§1.2; §2.1.1), rinnova una visione dualista e cartesiana in cui la mente e il mondo sono composte di due sostanze radicalmente diverse e in cui la comprensione o percezione del secondo da parte della prima deve necessariamente passare per un processo di rappresentazione interno della realtà esterna. Da qui, la lunga serie di riflessioni riguardo alle possibilità e alle modalità di rappresentazione del mondo da parte della mente hanno incluso l'intenzionalità sia come oggetto teorico da sottoporre a indagine per comprendere il collegamento tra mente e mondo, sia come strumento metodologico utile per sciogliere i nodi del processo di comprensione del funzionamento della mente umana (Fodor & Lepore 1992).

In secondo luogo, e in conseguenza alla prima affermazione, l'intenzionalità viene caratterizzata da Brentano come il *marchio del mentale*. Infatti, per Brentano l'intenzionalità è la proprietà distintiva tutti gli stati mentali e solo e soltanto di questi. In altre parole, l'intenzionalità in quanto proprietà è ciò che distingue i fenomeni psicologici da quelli fisici: a differenza di questi ultimi, come ad esempio la direzionalità di un oggetto nello spazio o una reazione chimica, stati mentali come credere, desiderare, amare, odiare, percepire, giudicare, ricordare, comprendere, sperare, ecc., sono sempre rivolti a qualcosa – un oggetto, una situazione, un concetto. Una pietra che cade non è diretta verso alcunché, mentre qualsiasi stato mentale è invece diretto verso un oggetto del mondo, che questo sia presente o meno al momento dell'atto di pensarlo, che sia reale o immaginario<sup>121</sup>.

Infine, l'intenzionalità è concepita come una relazione: relazione tra atti mentali e un contenuto, tra pensiero e oggetto. L'intenzionalità che lega, ad esempio, l'atto mentale di desiderare con l'oggetto o il contenuto desiderato costituisce una relazione in cui entrambi i *relata* hanno proprietà costitutive dell'atto intenzionale stesso. In quanto proprietà esclusiva degli atti mentali, la relazione intenzionale però non ha struttura causale, ma è, nei termini di Husserl, una relazione concettualmente primitiva

---

<sup>121</sup> La questione relativa alla possibilità di intenzionare, ovvero dirigere la propria attenzione mentale, verso oggetti di natura fantastica (come cerchi quadrati e unicorni) ha trovato numerose declinazioni nel pensiero di filosofi analitici (ad es. Meinong 1960) e scienziati cognitivi (Chrisolm 1973; Quine 1960). Per una rassegna sull'argomento si veda (Lyons 1995; Moran 2013).

(cfr. Taieb 2018: 2): ciò significa che, mentre nella relazione causale si identifica la necessità che uno dei *relata* contenga le condizioni di determinazione necessaria dell'altro, nella relazione intenzionale si riconosce che l'uno esibisce le condizioni di interpretazione possibile dell'altro (cfr. Mininni 2000). Tale differenza implica l'impossibilità di ridurre a rapporti di causa-effetto l'intenzionalità, ma apre alla possibilità di analizzare la struttura dell'intenzionalità in almeno due maniere, che hanno portato a una divisione di interpretazioni e alle formazioni di correnti di pensiero ben distinte nel panorama filosofico del '900. Come fatto notare fa Taieb (2018: 202):

The distinction between the intentional relation and the causal relation is closely connected with an account of mental acts from two different perspectives. The causal relation depends on the interaction between the environment and the subject (including its physiology). The account of this relation corresponds to the account of the *genesis* of the act: the act is the effect of the object and relates to that object as the cause of that act. The intentional relation, on the other hand, does not depend on anything other than the act itself; the account of this relation informs us about the *structure* of the act.

Entrambe le questioni hanno avuto un forte impatto sulla strutturazione del pensiero filosofico del Novecento e, in seguito, sulla elaborazione delle riflessioni riguardo la mente e la sua capacità di pensare oggetti del mondo sviluppata nelle scienze cognitive. Infatti, è a partire dalle differenti interpretazioni di queste due caratteristiche che le correnti analitica e fenomenologica della filosofia occidentale hanno costruito i propri approcci al concetto di intenzionalità. Entrambe queste correnti di pensiero fanno riferimento a Brentano come padre comune del concetto di intenzionalità e partono dalla sua definizione per costruire le proprie posizioni sulla caratteristica della mente e del pensiero di essere diretti verso qualcosa al di fuori di essi, ma elaborano riflessioni nettamente diverse nell'interpretazione del fenomeno (Lyons 1995). Da una parte, il concetto di intenzionalità diviene uno dei pilastri della fenomenologia, ripreso da uno dei più celebri allievi di Brentano, Edmund Husserl. Fu proprio Husserl a introdurre il concetto di intenzionalità all'interno del panorama filosofico a cavallo all'inizio del '900, ponendolo al centro della corrente fenomenologica e facendone il fondamento della sua metodologia di indagine della coscienza. Per Husserl, l'intenzionalità è la proprietà costitutiva della coscienza: non si può essere coscienti senza un oggetto che informi la coscienza su di sé, pertanto la direzionalità che la coscienza prende è costitutivamente implicata nel processo di formazione della coscienza stessa (Zahavi 2010; Gallagher & Zahavi 2008). In questo approccio, la divisione tra atti intenzionali ed enti intenzionali (l'oggetto dell'atto) è motivata da Husserl con una ragione metodologica: sono gli atti intenzionali a essere oggetto della fenomenologia, non gli enti (che possono essere investiti di atti anche molto diversi tra di loro) (Ales Bello 2005). Gli atti intenzionali tramite cui gli oggetti si offrono alla coscienza la definiscono come 'coscienza di', cioè le assegnano la sua natura intenzionale: in sintesi, l'essenza dell'intenzionalità consiste nel rendere possibili i fenomeni della coscienza quali processi di attribuzione di senso

(Duranti 1999). È importante sottolineare, come vedremo meglio in seguito, che il modo in cui si struttura in concetto di intenzionalità in fenomenologia a partire da Husserl (passando per Heidegger e Merleau-Ponty) rifiuta la definizione data da Brentano riguardo la incompatibilità di sostanza tra mente e mondo e, di conseguenza, la necessità di fare ricorso a strumenti concettuali come le rappresentazioni per metterle in comunicazione (Drummond 2012; Moran 2013; Zahavi 2005).

Dall'altra parte, nella filosofia analitica l'intenzionalità entra a far parte degli oggetti di indagine più investigati sotto una diversa dimensione. Quello che viene reso pertinente in questa sede non è tanto il modo in cui l'intenzionalità si dà all'esperienza del soggetto cosciente, quanto invece in rapporto al linguaggio e alla proprietà condivisa di riferirsi a oggetti del mondo esterno alla mente. In particolare, le riflessioni di Frege e Russell hanno contribuito a sviluppare una nozione di intenzionalità strettamente legata alle proprietà semantiche del linguaggio. Per il primo, il significato di un'espressione linguistica non è riducibile al referente esterno di cui parla, ma include anche la sua rappresentazione mentale: questo ha aperto la strada ad un'indagine dell'intenzionalità in termini di condizioni di verità e contenuto logico-semantic; per il secondo, la possibilità di dichiarare frasi il cui referente non esiste nel mondo esterno (ad es. "Il re di Francia è calvo") dimostra la possibilità di analizzare l'intenzionalità attraverso il modo in cui le descrizioni linguistiche realizzano il riferimento (cfr. Harney 2013: 37-71).

Domande al cuore della questione dell'intenzionalità, riguardo al suo ruolo nella struttura della mente e della cognizione o nei processi che strutturano la relazione tra mente e mondo, si pongono al centro delle scienze della mente fin dalla loro fondazione. Nelle scienze cognitive, l'intenzionalità è stata importata dal dominio filosofico come strumento per aprire una riflessione sul modo in cui i processi cognitivi possano rendere conto della relazione che si instaura tra il pensiero e il mondo, soprattutto grazie alle prime riflessioni sul tema sviluppate da Anscombe e Chisholm (Gozzano 1997; Moran 2013). Nonostante l'uniformità dei modi di considerare l'intenzionalità come elemento di partenza, il dominio delle scienze cognitive ha visto negli anni moltiplicare gli approcci relativi all'intenzionalità e al suo ruolo nella comprensione dei processi cognitivi. Per riuscire a comprendere il modo in cui l'intenzionalità è entrata anche nel dibattito degli studi sull'interazione tra umani e robot, possiamo fare appello a un testo ormai classico, che, ancora oggi, risulta al meglio descrivere il panorama delle teorie riguardo l'intenzionalità nelle scienze della mente.

#### *4.1.1 Tre approcci all'intenzionalità nelle scienze cognitive*

Il testo a cui facciamo riferimento è "The Intentionality All-Stars", scritto dal filosofo John Haugeland nel 1990. Declinando il campo delle scienze cognitive con una metafora basata sul baseball, Haugeland si pone il proposito di «outline and compare several of the main positions, and

to name some of their foremost defenders – to chose, in effect, an *All-Star Team* for the intentionality game» (1990: 386-387). In quegli anni, l'intenzionalità era uno degli argomenti più discussi all'interno delle scienze cognitive e della filosofia della mente, e Haugeland nota che le argomentazioni che nutrivano approcci differenti si riguardavano principalmente tre questioni fondamentali (1990: 384-385):

1. Il primo punto di interesse è relativo alla struttura relazionale dell'intenzionalità. Come visto, già a partire da Brentano questa si pone come una questione centrale per comprendere quella che viene definita la caratteristica identificativa degli atti mentali, soprattutto alla luce del fatto che la relazione non necessariamente deve prevedere l'*esistenza* di entrambi i *relata*. Già per i suoi allievi, come Husserl e Freud, la questione della in-esistenza (*in-existence*) dell'oggetto intenzionale aveva posto serie sfide interpretative riguardo cosa questo termine potesse significare: l'in-esistenza prevede che l'oggetto può anche non esistere nel mondo, come succede per unicorni o cerchi quadrati? Oppure va presa nel suo senso etimologico, dal latino *in-esse*, esistere all'interno (della mente)? In questo secondo caso, il termine denoterebbe le modalità di esistenza dell'oggetto intenzionale, ovvero esisterebbe nella mente nel momento stesso in cui è intenzionato, realizzandosi dunque in un'*in-esistenza intenzionale* (cfr. Jacob 2003).
2. Un'altra questione è relativa al carattere normativo dell'intenzionalità e al suo legame con la spiegazione causale. Frasi come “Credo che quel ristorante sia buono *perché* me l'ha consigliato Francesco, e Francesco non sbaglia mai sui ristoranti” o “Vado a mangiare in quel ristorante *perché* ho fame e niente mi impedisce di farlo” sono esempi di un doppio significato di *perché*, e quindi della relazione istituita tra le parti:

[...] the force of this “because” is also peculiar, in that it purports simultaneously to explain and to justify. The double purport is reflected in the double sense in the “reason” one does something. On the one hand, it must be *normative*, since my reasons, e.g., for coming to believe or acting, can always be criticised according how good or valid they are. On the other hand, it must be *causal*, since I can “have” a reason to do something, and do it, yet not do it “for” that reason - it's not the having of that reason that leads to my doing it. (Haugeland 1990: 384).

Porre l'accento su uno o sull'altro versante, ovvero su quello normativo o su quello causale, comporta delle differenze di pensiero riguardo a come l'intenzionalità può essere compresa e indagata, sia in relazione alla sua stessa struttura e a quella della mente, sia rispetto al rapporto che l'intenzionalità ha con il linguaggio e il contenuto semantico che veicola.

3. L'ultimo punto di discussione riguarda la differenza tra intenzionalità originale e derivata. La differenza tra un certo tipo di intenzionalità genuina e una “surrogata” era stata sollevata da alcuni filosofi e scienziati cognitivi fin dagli anni '70. Ad esempio, Searle (1980) distingue tra

intenzionalità intrinseca e relativa all'osservatore: in questa visione, gli stati mentali hanno intenzionalità intrinseca, ovvero in quanto conseguenza causata dalle proprietà fisiche e chimiche del cervello, mentre i segni esterni quali simboli, norme e istituzioni hanno l'intenzionalità derivata tipica del significato linguistico, la quale acquista valore solamente in relazione alla più genuina forma di intenzionalità intrinseca (cfr. anche Gozzano 1997: §4.4.3). Sulla stessa linea di pensiero e negli stessi anni, Block (1987) aveva introdotto la distinzione tra significato autonomo ed ereditato. Haugeland ci tiene a specificare che la sua distinzione tra intenzionalità *originale* e *derivativa* non debba essere confusa con questi, dal momento che la sua differenza non si poggia su un assunto metafisico circa l'origine dell'intenzionalità, ma sul fatto che, a prescindere da come qualcosa acquisti un'intenzionalità derivata (problema anche questo spinoso, come vedremo), ci debba essere necessariamente un'intenzionalità originale, ovvero non-derivativa. Per alcuni, come Searle, l'intenzionalità "originale" si dà a causa della organizzazione materiale del cervello, allo stesso modo in cui avviene la fotosintesi nelle piante (Searle 1983); per altri, invece, consiste in una proprietà intrinseca degli stati mentali rappresentazionali e della semantica del linguaggio (Fodor 1983); altri, come Richard Rorty o Jacques Derrida, credono che in nessun modo sia possibile trovare una intenzionalità originale né indagando i processi neurobiologici, né sondando i meccanismi logici interni al linguaggio: l'intenzionalità è solamente un altro degli strumenti socio-linguistico-culturali elaborati dall'essere umano per acquietare la propria sete di trascendentalismo (Haugeland 1990: 387)<sup>122</sup>.

Sebbene siano ormai passati più di tre decenni dalla sua pubblicazione, questo testo delinea ancora oggi una delle migliori riflessioni circa il dibattito sull'intenzionalità. Testimoni dell'attualità di questa visione sono i numerosi articoli che hanno ripreso recentemente la categorizzazione proposta da Haugeland e che ne hanno fatto il punto di partenza per analisi epistemologiche riguardo il concetto (Hutto & Satne 2015; Gallagher 2017; Gallagher & Miyahara 2012; Varga 2015; Veit 2022; Zawidzki 2015). Come è ovvio, tuttavia, negli ultimi trenta anni molti sono stati gli sviluppi sia sul tema dell'intenzionalità e la sua utilità nella spiegazione o nella descrizione di fenomeni mentali, sia riguardo le diverse interpretazioni della struttura dell'intenzionalità e delle possibilità di verificarne l'adeguatezza tramite esperimenti. Quello che faremo, dunque, sarà presentare gli approcci come originariamente raccolti da Haugeland, cercando allo stesso tempo di aggiornare le informazioni e

---

<sup>122</sup> A questa lista, Haugeland (1990: 388) aggiunge Skinner, il quale «is up against the wall defending the view that science can explain behavior without reference to intentionality, and therefore there is no reason to posit intentional states or entities». Questa posizione, come si evincerà dalla centralità acquisita dalla nozione di intenzionalità non solo nelle scienze cognitive in generale, ma anche negli studi di HRI, sembra essere l'unica posizione effettivamente invecchiata male. Per questo, abbiamo ritenuto di riportarla in nota.



fornendo così una panoramica per comprendere le diverse posizioni riguardo l'intenzionalità nelle scienze cognitive.

#### *4.1.1.1 Neo-Cartesianesimo: la visione internalista classica*

Il primo approccio individuato avente delle posizioni omogenee riguardo la questione dell'intenzionalità è quello Neo-Cartesiano. I maggiori sostenitori di questo approccio sono identificati da Haugeland (1990: 388-395) come successori di un approccio dualista della mente, idealmente o fattualmente eredi di Cartesio riguardo la divisione tra mente e mondo: le idee proposte da esponenti delle scienze cognitive classiche computazionaliste come Jerry Fodor e Zenon Pylyshyn rappresentano al meglio tale posizione, ma anche il filosofo John Searle viene incluso in questo gruppo, dal momento che fornisce un'interpretazione dell'intenzionalità come proprietà esclusiva degli stati mentali interni al sistema cognitivo e pone le rappresentazioni come ponte che garantisce comunicabilità tra mente e mondo.

L'idea alla base di questo approccio è la seguente: l'intenzionalità è la caratteristica degli stati mentali (interni alla mente) di essere circa o rivolti a qualcosa nel mondo (esterno alla mente). Il modo attraverso cui tale direzionalità o attinenza possa avvenire e quale sia la sua struttura sono gli aspetti principalmente indagati all'interno di questo approccio. Non si farà fatica a riconoscere nella definizione proposta una forte somiglianza alla questione dell'intenzionalità per come originariamente posta da Brentano: i termini implicati nell'indagine riguardano principalmente la natura del contenuto intenzionale, il suo rapporto con il mondo e le modalità di costruzione dell'oggetto intenzionale. Non a caso, Gallagher (2017: 66) assimila l'approccio neo-cartesiano per come proposto da Haugeland a una visione Brentaniana dell'intenzionalità e riassume in poche parole il cuore della posizione: «the mind is in some way discontinuous with everything around it, but at the same time, naturalistically continuous with the brain». In sostanza, l'intenzionalità è localizzabile nel cranio dell'individuo, proprietà intrinseca dei suoi stati mentali interni e distintiva della mente.

Per esplicitare ulteriormente gli assunti alla base dell'approccio neo-cartesiano, Horgan e Kriegel (2008) espongono in sei proposizioni i capisaldi di questo approccio:

1. La mente esibisce intenzionalità in virtù della sua natura di coscienza fenomenica (c'è quindi una intenzionalità intrinseca ai fenomeni di coscienza).
2. Il contenuto intenzionale è una proprietà interna degli stati mentali e può essere istanziato in “cervelli nella vasca”<sup>123</sup>, dal momento che l'ipotesi di fondo è che ad ogni stato mentale corrisponda in modo univoco una serie di correlati neurali.

---

<sup>123</sup> Il “cervello nella vasca” è un esperimento mentale elaborato dal filosofo della mente Hilary Putnam (1981) ispirandosi a un saggio di Daniel Dennett (1978c). L'esperimento mentale prevede che un cervello venga rimosso dal corpo di un individuo e immerso in una vasca in cui all'interno ci sono sostanze nutritive che riescono a tenerlo in vita

3. L'intenzionalità è soggettiva, ovvero l'accesso alla sua struttura è data solamente al soggetto che la esperisce.
4. L'intenzionalità è propria sia degli stati sensoriali, sia di quelli cognitivi.
5. L'intenzionalità è non-derivativa, ovvero ha origine dagli stati mentali interni.
6. L'intenzionalità è il marchio del mentale.

Se, dunque, i) la mente è qualcosa strettamente confinata all'interno dell'individuo, ii) i contenuti intenzionali fanno parte della mente e iii) gli oggetti che vengono pensati sono al di fuori di questa nel mondo esterno, è necessario qualcosa che possa permettere la comunicazione tra queste due sfere di pertinenza. Il ruolo delle rappresentazioni mentali diventa centrale in questo approccio proprio perché permette a sostanze diverse, in pieno stile cartesiano, di entrare in comunicazione: la mente e il mondo sono definitivamente scissi e l'unico strumento utile per collegarli e farli comunicare sono le rappresentazioni, le quali garantiscono all'oggetto intenzionale di acquisire un contenuto semantico esprimibile nella sintassi del linguaggio.

Si può tranquillamente affermare che questa è, ancora oggi, la visione più diffusa nelle scienze cognitive riguardo l'intenzionalità e gli aspetti che devono essere indagati. Infatti, se prendiamo ad esempio la definizione di 'Intenzionalità' all'interno della *Stanford Encyclopedia*, vedremo che questa è definita come «the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs. To say of an individual's mental states that they have intentionality is to say that they are mental representations or that they have contents» (Jacob 2003). In sintesi, in questo approccio l'intenzionalità è la proprietà degli stati mentali di essere diretti verso qualcosa del mondo, e può essere realizzata solamente attraverso modalità di rappresentazione che permettono al mondo esterno di acquistare significato, in qualche maniera, da parte della mente.

Esempi concreti di questa visione si possono ritrovare, come abbiamo anticipato, negli approcci classici delle scienze cognitive, come quello proposto da Jerry Fodor. Per Fodor (1975, 2008), l'intenzionalità è una proprietà fondamentale di stati mentali quali credenze, desideri, percezioni ed emozioni, che consente loro di acquisire un contenuto semantico, ovvero un significato descrivibile in termini proposizionali. Nella sua visione, Fodor difende una visione *realista* dell'intenzionalità, in cui a ogni dimostrazione di comportamento intenzionale da parte dell'agente deve necessariamente corrispondere uno stato interno del sistema riconoscibile a sua volta come intenzionale – non

---

anche al di fuori del corpo. Collegati al cervello nella vasca ci sono degli elettrodi che conducono input capaci di far credere al cervello stesso di essere ancora all'interno del corpo, simulando gli stati sensoriali che avrebbe davanti a esperienze vive, esattamente come se si trovasse ancora all'interno di un individuo. In questo scenario, non ci sarebbe modo per il cervello di sapere se quella che sta vivendo sia un'esperienza genuina o simulata: tale dubbio apre ad una serie di discussioni riguardo alla veridicità dell'esperienza cosciente, al solipsismo e allo scetticismo radicale, argomenti tipici – anche se aggiornati – della corrente di pensiero che si rifà a Cartesio, in particolare alla sua argomentazione sul genio maligno.

necessariamente individuabile come configurazione fisica del sistema (livello dell'*hardware*), ma rintracciabile come correlato computazionale (livello del *software*). Nella visione della cognizione come computazione proposta da Fodor,

following R requires that one's behavior have a certain kind of etiology; roughly, that one's intention that one's behavior should conform with R explains (or partly explains) why it does conform to R. A fortiori, you aren't following R unless R is the 'intentional object' of one of your mental states. (Fodor 1975: 37-38)

L'ipotesi del linguaggio del pensiero (*Language of Thought*, o mentalese) risulta fondamentale in questo quadro, poiché stabilisce i criteri minimi per l'acquisizione di un contenuto intenzionale da parte di un oggetto e perché specifica la struttura logica del meccanismo intenzionale, ovvero quella simbolico-sintattica propria del linguaggio. Ma una caratteristica fondamentale non deve essere tralasciata: il sistema deve essere in grado di manipolare i simboli in maniera adeguata e deve saper garantire un determinato grado di capacità computazionale, pena l'esclusione dal novero dei sistemi aventi intenzionalità originale. Infatti, come sottolineato da Haugeland (1990: 394), in questa visione un sistema non può avere intenzionalità originale o genuina a meno che possa avere pensieri circa sé stesso e il suo stesso pensiero, o riguardo i pensieri degli altri, la differenza fra verità e falsità, norme e valori, vita e morte, ecc.:

[...] it seems clear that overly simple-minded systems are not "minded" at all – even if they satisfy all the above characterizations. Arguably, for instance, the symbols in most ordinary computers, including sophisticated "expert systems," are only derivatively intentional, even though they are semantically articulated [...].

L'ipotesi ha fatto breccia nelle scienze cognitive, insieme all'impostazione computazionalista della mente, fino a costituirsi come il paradigma dominante riguardo il fenomeno dell'intenzionalità. Contemporanee evidenze della attualità del pensiero neo-cartesiano non si hanno solamente nelle scienze cognitive che ancora oggi mantengono una visione computazionalista della mente e della cognizione (Milkowski 2013; Colombo & Piccinini 2023), ma anche in alcuni degli approcci che avanzano una visione più incarnata della mente. Esempio d'eccellenza è la teoria della mente estesa (Clark & Chalmers 1998; Clark 2008) che, nonostante mostri significative differenze in merito alle idee su mente e cognizione, esibisce alcuni tratti in comune con l'idea neo-cartesiana di intenzionalità: per entrambi gli approcci, l'intenzionalità è qualcosa che può e deve avere riscontro nelle modalità di dispiegamento dei meccanismi cognitivi, fattore che comporta la possibilità di indagarne la struttura materiale in un paradigma scientifico; allo stesso tempo, entrambi concordano sulla necessità di rappresentazioni mentali che fungono da strumento cognitivo indispensabile per l'acquisizione del significato. In sintesi, a fronte di una radicale differenza circa la localizzazione della mente all'interno dell'individuo (come per Fodor) o nel contesto distribuito della dinamica tra corpo, cervello e oggetti del mondo (come per Clark), l'intenzionalità è una caratteristica *interna* alla mente, che la costituisce

dal di dentro e che si dispiega in relazione alle capacità di elaborazione semantica che definiscono il sistema.

Analizzeremo in seguito (§4.1.2) le declinazioni di questo approccio nella cognizione sociale. Per ora, ciò che ci preme sottolineare è la visione internalista e solipsista dell'approccio neo-cartesiano all'intenzionalità. Questi tratti si affievoliscono in modo incrementale in approcci alternativi a quello neo-cartesiano, a cominciare da quello neo-comportamentista.

#### *4.1.1.2 Neo-Comportamentismo: strategie di previsione*

Il secondo approccio elencato da Haugeland (1990: 395-403) è quello neo-comportamentista, i cui antecedenti filosofici si rintracciano in personalità come Gilbert Ryle e Ludwig Wittgenstein. Come suggerisce il nome, tuttavia, i debiti teorici messi in risalto fanno riferimento alla corrente psicologica del comportamentismo: da quest'ultimo, questo approccio all'intenzionalità eredita l'idea che il sistema sia in sostanza una scatola nera, una *black box* il cui funzionamento interno è impossibile da svelare, o quanto meno non immediatamente dato all'osservatore. Di conseguenza, data l'impossibilità di indagare il meccanismo intenzionale per come si genera all'interno della mente, ciò su cui questi approcci si focalizzano è il comportamento del sistema, ovvero sui «pattern exhibited in perceptions and actions» (Haugeland 1990: 396).

Il primo aspetto sottolineato da Haugeland è la discordanza con l'approccio neo-cartesiano: mentre per il primo l'intenzionalità è intrinseca agli stati mentali del sistema e localizzabile nella sua mente, per questo approccio l'intenzionalità è qualcosa che viene ascritta o attribuita al sistema cognitivo nel suo complesso. In particolare, tre sono le differenze con l'approccio neo-cartesiano che definiscono a loro volta l'identità di quello neo-comportamentista:

- i. L'indagine dell'intenzionalità non si poggia su un assunto metafisico che la localizza all'interno della mente dell'individuo. Al contrario, il modo attraverso cui un sistema può essere considerato avere intenzionalità per i neo-comportamentisti è dimostrando «a pattern in the interactions between the system and its environment» (Haugeland 1990: 395).
- ii. Questo pattern di interazioni non è concepito come una serie di stimoli fisici e risposte cognitive, ma come sistemi di percezioni e azioni (*Ibidem*). In questa visione, gli stati cognitivi del sistema acquistano intenzionalità in quanto collegati causalmente alle percezioni e alle azioni realizzate dal sistema, ovvero gli stati cognitivi possono essere detti intenzionali se le loro relazioni causali con le percezioni e le azioni sono tali da fare senso come contenuti rappresentazionali significativi per un comportamento efficace.

For states of a system to be intentional, the causal relations connecting them to one another and to certain environmental interactions must account for the fact that the latter consistently make sense as contentful perceptions and actions. It is an empirical question

whether unarticulated states of material system can indeed be so causally structured (that is, whether perception and action are in fact possible without a language of thought); but, if they can, there is no conceptual barrier to ascribing them as cognitive and rational. (Haugeland 1990: 398)

Per comprendere meglio il concetto, si pensi alla situazione in cui si trova un cacciatore che ha appena visto una lepre correre dietro a un cespuglio nel bosco. Secondo l'approccio neo-comportamentista, questa situazione può essere descritta in termini di percezioni e azioni: il cacciatore vede (percezione) che il cespuglio si muove e ciò lo porta ad avvicinarsi al cespuglio (azione) per agguantare la lepre. Allo stesso tempo, gli stati cognitivi del cacciatore sono collegati alla situazione esterna appena descritta: il cacciatore ha la credenza che ci sia una lepre nascosta dietro a quel cespuglio e ha il desiderio di catturarla. Sulla base di questi stati mentali, il cacciatore elabora di conseguenza una strategia affinché possa incontrare i propri desideri: "mi avvicinerò lentamente per non spaventarla e poi mirerò in modo preciso", penserà. Per i neo-comportamentisti, gli stati cognitivi del cacciatore possono essere intenzionali perché integrati in un sistema in cui le percezioni e le azioni sono collegate tra di loro e significative, rappresentate adeguatamente all'interno della sua mente: le percezioni sono interpretate come indizi, i quali nutrono le credenze, che a loro volta orientano la strategia, la quale è realizzata tramite l'azione.

- iii. Di conseguenza, gli stati mentali sono considerati intenzionali quando possono essere compresi come parte di un sistema in cui le percezioni e le azioni sono integrate in un quadro strategico e comportamentale più ampio. Gli stati mentali non vengono visti come stati computazionali del sistema, non si danno come calcoli logici o elaborazioni interne ma come *disposizioni al comportamento*, ovvero come modalità di risposta a specifiche situazioni, come motivazioni o predisposizioni all'assunzione di determinate strategie di azione e spiegazione degli eventi. L'intenzionalità è quindi ciò che garantisce la tenuta del sistema, ovvero quel legame attraverso cui si stabilisce la struttura causale delle disposizioni e credenze sulle azioni osservabili. In questo aspetto si scorge la ragione per cui l'approccio può essere considerato rifarsi al comportamentismo e allo stesso tempo discostarsene: mentre i comportamentisti classici come Skinner tendevano a minimizzare il ruolo degli stati mentali interni, concentrandosi quasi esclusivamente sui comportamenti osservabili e misurabili, la nuova versione attribuisce loro un ruolo fondamentale nella comparsa dell'intenzionalità nel comportamento, anche se questo rimane l'oggetto di studio privilegiato.

Seguendo la critica radicale di Ryle (1949) al dogma del *ghost in the machine*<sup>124</sup>, l'intenzionalità non è localizzata in una dimensione interna e inaccessibile della mente, ma è interamente pubblica e osservabile attraverso il comportamento: come l'università non trova la sua identità in nessuno dei luoghi che la compongono, così la mente non deve essere trattata in quanto oggetto, entità o sostanza distinta dal corpo, quanto invece riconosciuta come manifestazione di disposizioni e comportamenti. La posizione di Wittgenstein (1953)<sup>125</sup> ci consente di sviluppare ulteriormente questa idea, dal momento che sosteneva che gli stati mentali non esistono in isolamento dalla prassi sociale in cui si manifestano. In altre parole, l'intenzionalità non è un'entità che giace nascosta nella mente, ma qualcosa che si manifesta visibilmente nelle azioni dell'individuo nel contesto sociale. Da questa prospettiva, il comportamento stesso è già intenzionale, senza bisogno di inferire stati mentali interni.

Tali assunti non solo descrivono un approccio molto diverso da quello classico all'intenzionalità, ma comportano anche una conseguenza sul piano delle considerazioni che ne derivano. Infatti, l'attribuzione di intenzionalità all'agente non deve andare di pari passo con una specifica struttura interna degli stati mentali, né con una necessaria e sufficiente capacità computazionale posseduta dallo stesso che dovrebbe garantire la genuinità della proprietà rilevata. Le ascrizioni non sono traduzioni di cosa succede all'interno della mente dell'agente, ma modalità adottate per garantire la prevedibilità dello stesso attraverso l'elaborazione di strategie efficaci: «intentionality is manifest in behavioral tokens not as “reasonable sense” of a text, but as the *competence* of an agent» (Haugeland 1990: 398). Per questo, secondo questo approccio il valore normativo dell'intenzionalità non è dato dai valori di verità, ma dal successo che dimostra il comportamento intenzionale, e le relazioni logiche tra gli stati ascritti al sistema non sono primariamente semantiche e ontologiche, ma strumentali, ovvero orientati come mezzi e fini.

---

<sup>124</sup> Il dogma del *ghost in the machine* è al centro della critica proposta da Ryle al dualismo cartesiano mente-corpo. In particolare, Ryle descrive il dogma svilupparsi alla separazione della mente (the ghost) dal corpo (la macchina). La critica di Ryle si sviluppa a partire dalla definizione di “categorical mistake” (Ryle 1949: 5-8), che viene reso chiaro attraverso l'esempio, ormai diventato celebre, sull'impossibilità di localizzare l'università come istituzione in uno o più edifici (cfr. Tanney 2022).

<sup>125</sup> L'idea di Wittgenstein riguardo l'impossibilità che gli stati mentali siano riducibili a entità interne alla mente è sintetizzata nel famoso passo del *coleottero*: «Se dico di me stesso che soltanto dalla mia personale esperienza io so che cosa significa la parola «dolore», – non debbo dire la stessa cosa anche agli altri? E come posso generalizzare quest'unico caso in maniera così irresponsabile? “Ora qualcuno mi dice di sapere che cosa siano i dolori soltanto da se stesso! Supponiamo che ciascuno abbia una scatola in cui c'è qualcosa che noi chiamiamo «coleottero». Nessuno può guardare nella scatola dell'altro; e ognuno dice di sapere che cos'è un coleottero soltanto guardando il suo coleottero. – Ma potrebbe ben darsi che ciascuno abbia nella sua scatola una cosa diversa. Si potrebbe addirittura immaginare che questa cosa mutasse continuamente. – Ma supponiamo che la parola «coleottero» avesse tuttavia un uso per queste persone! – Allora non sarebbe quello della designazione di una cosa. La cosa contenuta nella scatola non fa parte in nessun caso del giuoco linguistico; nemmeno come un qualcosa: infatti la scatola potrebbe anche essere vuota. – No, si può ‘dividere per’ la cosa che è nella scatola; di qualunque cosa si tratti, si annulla.

Questo vuol dire: Se si costruisce la grammatica dell'espressione di una sensazione secondo il modello ‘oggetto e designazione’, allora l'oggetto viene escluso dalla considerazione, come qualcosa di irrilevante» (Wittgenstein 1953: par.293).

Nelle scienze cognitive, il campione di questa prospettiva è stato il filosofo Daniel Dennett, il quale ha contribuito in maniera sostanziale allo sviluppo di questa corrente di pensiero sull'intenzionalità, definendone gli assunti ed elaborando una finissima teoria denominata *atteggiamento intenzionale* (o *intentional stance*; Dennett 1971, 1987, 1988). Nell'approccio di Dennett, l'intenzionalità non è qualcosa che si rintraccia nella struttura interna della mente ma, appunto, qualcosa che viene *ascritta* o *attribuita* a un sistema che mostra un comportamento passibile di essere descritto attraverso il suddetto atteggiamento. In altre parole, l'intenzionalità per Dennett si dà in quanto *strategia* di predizione del sistema che ci troviamo davanti, osservando il modo in cui si comporta in una determinata situazione.

Come abbiamo visto nell'esempio del cacciatore e della lepre, la strategia che il cacciatore sviluppa per raggiungere i propri obiettivi e soddisfare i propri desideri si basa, da una parte, sulle sue percezioni e, dall'altra, sulle sue credenze circa gli stati del mondo, l'efficacia delle azioni intraprese e gli effetti possibili che la strategia in questione avrebbe se ottimamente realizzata. Ma come arriviamo ad avere la capacità di mettere in relazione percezioni, desideri e obiettivi e a costituirci delle credenze che direzionano l'elaborazione della strategia di comportamento?

Nella teoria di Dennett, un ruolo fondamentale è ricoperto dalla psicologia del senso comune, o *folk psychology*, ovvero quella particolare abilità cognitiva che l'essere umano ha di comprendere ciò che avviene, «our everyday way of understanding, or rationalizing, intentional actions in mentalistic terms» (Hutto & Ravenscroft 2021). Il principio su cui si poggia la teoria dell'atteggiamento intenzionale consiste nell'ipotizzare che la *folk psychology*, il nostro uso quotidiano del linguaggio mentalistico, sia lo strumento attraverso cui le credenze<sup>126</sup> – e tutti gli stati mentali – che direzionano l'elaborazione delle nostre strategie si costituiscono<sup>127</sup>. Nelle parole di Dennett,

Suggerisco che la psicologia del senso comune sarebbe meglio interpretabile come un calcolo razionalistico di interpretazione e previsione – un metodo interpretativo idealizzante, astratto, strumentale, che si è evoluto perché funziona, e che funziona perché noi ci siamo evoluti. Noi entriamo in rapporto reciproco come *sistemi intenzionali*, cioè come entità il cui comportamento è prevedibile sulla base del metodo che consiste nell'attribuire credenze, desideri e acume razionale, conformemente ai principi elementari che seguono:

- 1) Le credenze di un sistema sono quelle che esso *dovrebbe avere*, date le sue capacità percettive, i suoi bisogni epistemici e la sua biografia. [...]

---

<sup>126</sup> «Che cosa sono le credenze? La psicologia del senso comune, o folk psychology, sostiene che le *credenze* sono stati di persone, che veicolano informazione, che nascono dalle percezioni e conducono, insieme a *desideri* appropriatamente connessi, a un'azione intelligente.» (Dennett 1987: 71)

<sup>127</sup> «Noi utilizziamo costantemente la psicologia del senso comune, per spiegare e prevedere il nostro comportamento reciproco; ci attribuiamo a vicenda credenze e desideri senza esitazioni – e in modo del tutto inconsapevole – e passiamo una parte essenziale della nostra esistenza da svegli a esprimere il mondo – senza escludere noi stessi – in questi termini.» (Dennett 1987: 73)

- 2) I desideri di un sistema sono quelli che esso *dovrebbe avere*, dati i suoi bisogni biologici e i mezzi più praticabili per soddisfarli. [...] desideri “anomali” sono attribuibili se si possono raccontare storie particolari.
- 3) Il comportamento di un sistema consisterà di quegli atti che per un agente con quelle credenze e con quei desideri *sarebbe razionale* compiere. (Dennett 1987: 74-75, corsivi dell'autore)

Uno degli aspetti più interessanti che possiamo rilevare in questo passaggio è relativo al ruolo che la dimensione condizionale gioca nell'adozione dell'atteggiamento: se il sistema *x* fosse in grado di avere stati mentali intenzionali, allora si comporterebbe nel modo *y* in una data situazione<sup>128</sup>.

Ora, abbiamo visto che il modo in cui la strategia di previsione e interpretazione del sistema consiste nell'individuare il legame causale tra gli stati percettivi del sistema stesso, le sue credenze riguardo a stati del mondo e le sue strategie di comportamento, ma non tutte le strategie di predizione e interpretazione possono essere basate sulla stessa logica. Predire il comportamento di un ombrello, che mi riparerà dalla pioggia quando lo aprirò sopra alla mia testa, e quello del mio amico Flavio, che lascia lo sportello della lavatrice chiuso quando invece dovrebbe rimanere aperto, sembra non basarsi sulle stesse logiche. Se l'ombrello non riesce a ripararmi dalla pioggia posso dire che è rotto, ma dirlo del mio amico Flavio che continua a chiudere lo sportello della lavatrice non sembra avere lo stesso grado di efficacia interpretativa. Come può, quindi, tale approccio riuscire a dare conto della differenza che c'è tra l'interpretazione del comportamento di un essere umano, quella di un ombrello che si rompe e quella di una pietra che cade da un burrone? Tra prevedere quali saranno le prossime azioni realizzate dal cacciatore e quelle che potrebbero non far funzionare il fucile quando è il momento di sparare?

Per giustificare tale differenza, Dennett elabora una gerarchia di atteggiamenti che l'essere umano adotta in corrispondenza di sistemi differenti. Ognuno degli atteggiamenti elencati consiste nell'adozione di una logica diversa per l'interpretazione del sistema, il quale viene spiegato in base al livello di complessità del comportamento che manifesta e alla natura dei suoi stati mentali.

Il primo livello, il più “elementare”, è l'atteggiamento fisico, o *physical stance*. L'atteggiamento fisico è la strategia più concreta, basata sulla conoscenza della fisica elementare che governa la materialità del mondo. Secondo Dennett, è l'atteggiamento attraverso cui si interpreta la mela che cade dall'albero, attraverso cui si prevede il punto in cui atterrerà una palla lanciata in aria, ma anche il moto degli astri e il fatto che la sedia reggerà il mio peso quando la userò per accomodarmi alla scrivania.

---

<sup>128</sup> «Come attribuiamo i desideri (preferenze, obiettivi, interessi) sulla cui base formuleremo poi l'elenco delle credenze? Noi attribuiamo i desideri che il sistema *dovrebbe avere*». (Dennett 1987: 36)



Consideriamo la strategia fisica o atteggiamento fisico; se si vuol prevedere il comportamento di un sistema, bisogna determinare la sua costituzione fisica (scendendo quasi sempre a livello microfisico) e la natura fisica delle spinte cui è sottoposto, e usare la nostra conoscenza delle leggi fisiche per prevedere gli esiti per ogni *input*. (Dennett 1987: 31)

Non sempre, però, il funzionamento di un'entità fisica può essere interpretato basandosi esclusivamente sulle leggi che lo governano, dal momento che potrebbero non essere sufficienti per raggiungere il grado di soddisfazione necessario, o potremmo non conoscere quelle giuste. In questi casi, Dennett parla di un cambiamento nell'atteggiamento, che passa dal fisico al progettuale, o *design stance*. Secondo questo atteggiamento, il sistema deve essere interpretato secondo le logiche di costruzione e funzionamento che il sistema incorpora. Esempi canonici dei sistemi interpretabili assumendo tale atteggiamento sono gli elettrodomestici – se il frigorifero è rotto e non si *comporta* come dovrebbe, ovvero se non raffredda, l'interpretazione migliore fa appello alle modalità di funzionamento e progettazione dell'oggetto: si sarà rotto qualcosa.

Talvolta, dato un evento qualsiasi, è più efficace passare immediatamente dall'atteggiamento fisico a quello che definisco atteggiamento progettuale, ove si prescinde dagli effettivi (probabilmente confusi) dettagli della costituzione fisica di un oggetto e, sulla base dell'assunto che esso ha una certa destinazione, si prevede che si comporterà così come *è stato progettato che si comporti* in diverse circostanze. (Dennett 1987: 32)

Lo stesso vale per l'atteggiamento che si adotta di fronte ad automobili, strumenti musicali, macchine del caffè e dispositivi che incorporano intelligenze artificiali. È proprio pensando a questa ultima classe di oggetti, come i computer, che Dennett descrive il funzionamento dell'atteggiamento progettuale, giustificando lo scarto tra *physical* e *design stance*:

[...] la maggior parte di coloro che usano i computers non hanno la più pallida idea di quali principi fisici siano responsabili del comportamento altamente affidabile, e quindi prevedibile, del computer. Ma se essi hanno una chiara idea di ciò che il computer è destinato a fare (il che comporta una descrizione delle sue operazioni per ognuno dei molteplici livelli d'astrazione possibili), allora sono in grado di prevedere il suo comportamento con grande precisione e attendibilità, previsione suscettibile di sconfirma solo in casi di cattivo funzionamento fisico del computer stesso. (Dennett 1987: 32)

Quello che si prevede, dunque, adottando questo atteggiamento è solamente il comportamento programmato di un sistema. Tuttavia, anche in questo caso, non sempre si possiede la conoscenza delle leggi che governano il funzionamento del sistema come diretto da uno schema, ed è qui che si fa riferimento alla forma più astratta di atteggiamento: quello intenzionale, o *intentional stance*.

Talvolta anche l'atteggiamento progettuale è praticamente inaccessibile, e allora esiste, nondimeno, un altro atteggiamento o strategia da adottare: l'atteggiamento intenzionale. Esso funziona nel modo seguente: innanzitutto si stabilisce di trattare l'oggetto di cui si deve prevedere il comportamento come un agente razionale; poi si individuano le credenze che l'agente in questione dovrebbe avere, data la sua posizione e il suo scopo nel mondo. Quindi si determinano i desideri che dovrebbe avere, sulla base delle stesse considerazioni, e infine si può prevedere che quest'agente razionale agirà per favorire i suoi obiettivi alla

luce delle sue credenze. Un piccolo ragionamento pratico sulla base dell'insieme di credenze e desideri prescelto produrrà in molti casi – ma non in tutti – una decisione su ciò che l'agente dovrebbe fare; cioè una previsione su ciò che l'agente *farà*. (Dennett 1987: 33)

In questo quadro, dunque, l'atteggiamento intenzionale è solamente una delle strategie possibili, che si adotta quando il sistema è suscettibile di essere descritto nei termini mentalistici della *folk psychology*: «l'atteggiamento intenzionale è infatti una macchina per generare o progettare circostanze aneddotiche – stratagemmi, tranelli e altre prove del fuoco intenzionalistiche – e prevederne gli esiti» (Dennett 1987: 337). Ma non tutte le strategie sono buone. Si pensi, ad esempio, a una strategia largamente diffusa, e dalle origini antichissime, per prevedere il comportamento di una persona: prendere come riferimento la data, l'ora e il luogo della sua nascita, individuare la costellazione più luminosa nel cielo in quel preciso momento e correlare ognuna di queste con delle caratteristiche caratteriali specifiche. Sebbene la strategia sia complessa, conoscere l'oroscopo di una persona difficilmente ci darà informazioni certe sulla sua natura di agente intenzionale e sulla struttura delle sue credenze in quanto disposizioni al comportamento. Sarà dura per qualcuno da accettare, ma l'oroscopo dei segni zodiacali non ha alcuna validità per la previsione del comportamento umano, e non solo perché potrebbe essere considerata un esempio di *design stance* di un progettista superiore.

Anche definire il mio amico Flavio “rotto” quando chiude lo sportello della lavatrice è altrettanto insoddisfacente. In questo caso, potrei fare appello a un significato metaforico del termine, ma la strategia più efficace sarebbe senza dubbio pensare che Flavio sia sbadato, che non mi abbia ascoltato, o che magari abbia voluto farmi un dispetto. Nel primo caso, richiamare la rottura del sistema-Flavio consiste nell'adottare un atteggiamento progettuale, che non risulta però molto efficace; ragionare in termini dei suoi potenziali stati percettivi e mentali, invece, mi pone nella situazione di poter elaborare una strategia più efficace in futuro per realizzare il mio desiderio (di farmi ascoltare).

Questa storia ci fa meglio comprendere quali siano, secondo i neo-comportamentisti, i criteri a cui ci riferiamo per comprendere l'adeguatezza dell'atteggiamento adottato per interpretare il sistema. L'approccio neo-comportamentista definisce come strumenti di validazione della strategia intenzionale l'adesione delle previsioni al comportamento del sistema: in poche parole, se la strategia ha successo nel predire il modo in cui il sistema si comporta, allora il sistema può dirsi intenzionale, e tanto basta. Non c'è bisogno di alcun correlato ontologico all'interno dell'individuo per far sì che l'ascrizione di intenzionalità avvenga o sia giustificata, ma solamente che l'atteggiamento intenzionale si dimostri i) motivato da comportamenti messi in atto dal sistema all'interno dei suoi pattern di comportamento e ii) una tattica di predizione adeguata ed efficace, anche se a volte esagerata per predire il comportamento dell'agente.

In riferimento a questo secondo punto, c'è un'ultima caratteristica fondamentale da sottolineare riguardo al concetto neo-comportamentista di intenzionalità, che riguarda la tendenza rilevata sia da

Dennett, sia da Haugeland, di adottare l'atteggiamento che, *per eccesso*, consente di interpretare il sistema osservato. Nelle parole di Haugeland (1990: 400), questo rivela una tensione nell'adozione degli atteggiamenti, i quali potrebbero portarci di volta in volta a sopravvalutare le capacità del sistema che ci troviamo a interpretare, seguendo un principio di massimizzazione della competenza esibita dal sistema<sup>129</sup>. È in questi casi che, se viene adottato l'atteggiamento intenzionale nei confronti di un oggetto realizzato secondo una progettualità, per Dennett si parla di antropomorfismo:

Per la gente che ha scarsa propensione per la meccanica, l'interpretazione intenzionale di un semplice termostato costituisce un supporto utile e in larga misura innocuo, ma quelli che tra noi sono ingegneri possono comprendere appieno il suo funzionamento interno, senza ricorrere a questa antropomorfizzazione. Può darsi sia vero che gli ingegneri più abili trovino praticamente impossibile conservare una chiara concezione di sistemi più complessi [...] senza cadere in un atteggiamento intenzionale [...], ma questo è solo un caso più sviluppato della fragilità epistemica umana. (Dennett 1987: 42)

In sostanza, la capacità del sistema di manifestare degli indizi minimi di comportamento intenzionale è condizione sufficiente affinché l'individuo adotti l'*intentional stance* nei confronti del sistema osservato: se la strategia è funzionale nel predire il comportamento del sistema, non importa quali siano i motivi, siano essi giusti – ovvero sia riscontrabile una corrispondenza con gli stati intenzionali interni del sistema – o sbagliati (perché magari il sistema non ha stati intenzionali interni).

Fondamentalmente, questa gerarchia delinea un percorso in cui la spiegazione del comportamento dell'altro (entità o individui) viene concepita in modo sempre più astratto e guidato dall'intenzionalità. Non viene data importanza al livello ontologico dell'intenzionalità, ossia non importa se l'agente sia o meno un vero "credente" (*true believer*, Dennett 1997); l'unica cosa rilevante è che il suo comportamento possa essere predetto o compreso in termini di attribuzione di stati mentali, e non semplicemente adottando spiegazioni meccanicistiche o fisiche.

Vedremo in seguito come questo approccio sia, almeno formalmente, quello maggiormente adottato all'interno degli studi di HRI per comprendere come l'intenzionalità viene attribuita al robot (§4.2). Ciò che vale la pena evidenziare in questo momento è il fatto che questo approccio si basa su una razionalità estrema: secondo i neo-comportamentisti, comprendere gli altri, anche solo come agenti, significa ragionare sulle loro strategie e obiettivi che vengono estrapolati dal loro comportamento e scegliere quale degli atteggiamenti adottare per raggiungere la massima efficienza interpretativa.

---

<sup>129</sup> Haugeland (1990: 400) nota che lo stesso errore di mis-interpretazione del sistema può essere compiuto per difetto, cioè adottando un atteggiamento progettuale laddove, invece, sembri più adeguato uno intenzionale, ma Dennett non fa riferimento a questo caso come un errore.

#### 4.1.1.3 Neo-Pragmatismo: esternalismo e norme sociali

L'ultimo approccio presentato da Haugeland è quello neo-pragmatista (Haugeland 1990: 404-313), con esponenti come Martin Heidegger, Hubert Dreyfus e Robert Brandom<sup>130</sup>. Per questi autori, l'intenzionalità è una proprietà delle capacità cognitive degli agenti e dei loro stati mentali che viene acquisita grazie al fatto che l'essere umano è sempre immerso pragmaticamente in un mondo significativo fatto di norme, istituzioni, artefatti e legami sociali: solo in funzione dei ruoli che l'individuo assume in questa rete di pratiche, azioni, funzioni, oggetti e posizioni significativi, gli stati mentali possono acquisire un qualsiasi contenuto o significato.

Brandom descrive questo approccio all'intenzionalità in quanto dipendente da concetti e pratiche sociali e normativi attraverso il *deontic scorekeeping*, ovvero la pratica reciproca e implicita di tenere traccia delle azioni degli altri e delle nostre in termini di stato normativo (Brandom 1994). Secondo questa visione, elaborata principalmente nel contesto di una teoria del discorso (Brandom 2000), comprendiamo l'intenzionalità dell'altro in modo implicito in termini di impegni e capacità specificate dalle norme sociali, anche se non sempre riconosciamo esplicitamente tale stato normativo. Una descrizione chiara e sintetica di questo approccio è fornita da Gallagher (2017: 72):

[...] we track, and occasionally acknowledge, other's people intentionality in virtue of what they are doing and saying, what they are expected to do or say, what roles they play, what kind of place and time it is, and what such factors mean to us in the shared social situation, rather than by somehow looking for mental states hidden behind their behaviors. We ascribe intentionality to actions to the extent that we have a practical grasp on their socially instituted significance. (Gallagher 2017: 72)

L'esempio più diffuso per definire la nozione di *deontic scorekeeping* è la promessa. Mettiamo il caso che io prometta a Francesco di vederci domani a casa sua per giocare a scacchi. Quando io faccio la promessa, sto assumendo un impegno normativo: sto accettando la responsabilità di presentarmi domani nell'orario concordato a casa di Francesco. Nel caso non mi presentassi all'appuntamento, Francesco avrebbe tutto il diritto di telefonarmi per chiedermi dove sia; nel caso arrivassi a casa sua e mi mettessi a parlare della mia tesi, Francesco potrebbe volermi ricordare che gli accordi fossero altri e che la nostra partita di scacchi ci attende. In altre parole, Francesco tiene traccia (*deontic scorekeeping*) delle mie promesse e delle mie azioni e si aspetta che rispetti la parola data. Il riconoscimento sociale, in questo contesto, è fondamentale: non mantenendo la promessa di presentarmi nel luogo dell'appuntamento all'orario concordato andrei incontro a un giudizio sociale, presumibilmente non buono. Dovrei, peraltro, giustificare la mancata promessa, auspicabilmente in un modo che possa esimermi da un giudizio negativo: in sostanza, ogni promessa crea una rete di responsabilità e il *deontic scorekeeping* viene aggiornato in base al rispetto o alla violazione di tali

---

<sup>130</sup> L'approccio neo-pragmatista è abbracciato anche dallo stesso Haugeland (1990: 422, nota 22).

responsabilità. Questo processo si delinea come un'evoluzione dinamica che coinvolge non solo gli individui che si trovano nella rete, ma anche la comunità sociale che riconosce e valuta gli impegni.

In sostanza, l'intenzionalità secondo i neo-pragmatisti si origina dalle norme sociali e culturali che condividiamo con gli altri esseri umani e che informa la volontà dell'individuo di mantenere alto il *deontic scorekeeping* che lo definisce. Come riportato da Clapin (2002: 17), «mental properties are derived from social properties and not vice versa»: possiamo attribuire significato, e quindi contenuto, a un pensiero solamente nel contesto della vita condivisa in cui il comportamento del sistema aderisce o si *conforma* alle norme sociali.

Metaphysically, conformism works like the mechanism of inheritance and natural selection [...] it engenders a new *kind* of order [...] the order and structure in cultural institutions and practices is not only vastly greater than in any genetic or adaptive ethology, but of a new and different kind – made possible by the distinctive mechanism of social conformism [...] whole new categories of phenomena emerge, including [...] social norms and original intentionality. (Haugeland 1990: 407)

Di conseguenza, in questo quadro teorico l'ascrizione dell'intenzionalità ad un'entità è possibile solo se l'entità è capace di sostenere una serie di pratiche, scelte e impegni, e che questi siano declinabili attraverso pratiche sociali e linguistiche. Alcuni dei proponenti di tale corrente affermano che «only communities, not individuals, can be interpreted as having original intentionality [...]». The practices that institute the sort of normative status characteristic of intentional states must be social practices» (Brandom 1994: 61). In questo caso, dunque, la questione dell'intenzionalità originaria è chiara e esattamente contraria alle soluzioni sviluppate nell'approccio neo-cartesiano: mentre nel secondo l'intenzionalità origina dagli stati mentali dell'individuo poiché strettamente inerente alla loro conformazione interna, e si trasferisce a quella derivata delle istituzioni e delle pratiche sociali in un processo di attribuzione che va dal soggetto al mondo, nel caso dell'approccio neo-pragmatista è proprio il sistema delle pratiche socialmente condivise quello in cui risiede la possibilità di un'intenzionalità originale, e quella che attribuiamo alle persone e ai loro stati mentali non è che una forma derivata:

At third base, as at first base, the primary bearers of content are semantically articulated symbols, occurring in appropriate dynamic patterns, and [...] reducible to distinct and determinate formal tokens. Third base tokens, however, are external, not internal, to the system that use them. [...] At third base, the instituted intentionality of public symbols is *original intentionality*. The extant normative order in the communal pattern is *sui generis* and self-sustaining, via the mechanism of conformism; it is the fountainhead of all intentionality, public and private. Thus, insofar as this order is imposed on the behavior or states of individual community members in such a way as to confer intentionality also on them, that resulting private intentionality is *derivative*. (Haugeland 1990: 412, 413).

A sostenere la realizzazione del processo di sviluppo dell'intenzionalità a partire dalle norme condivise nella comunità interviene la nozione di pratica. Infatti, le istituzioni e le norme in questo

contesto non vanno viste come esclusivamente dipendenti dalla loro potenziale espressione nel linguaggio, ovvero rese esplicitamente in forma proposizionale. Al contrario, il concetto neo-pragmatista di intenzionalità abbraccia una concettualizzazione di 'norma' radicata nell'azione ed individuabile nelle disposizioni adottate dagli agenti. Le pratiche messe in atto dagli agenti, infatti, sono pregne di intenzionalità, nella misura in cui sono inserite in un contesto normativo sociale che ne determina la significatività. Anche la pratica discorsiva, quindi, il parlare, è considerata come azione che acquista significato nella misura in cui è inserita all'interno di un contesto di pratiche sociali condivise. Tale possibilità si estende anche a contesti in cui il linguaggio esplicito è assente, come nei casi di comunicazione non verbale o di azioni tacite che comunque veicolano significato: l'intenzionalità, nella prospettiva neo-pragmatista, non è legata esclusivamente alla rappresentazione proposizionale o alla deliberazione cosciente, ma può essere rintracciata anche in azioni e comportamenti impliciti, purché questi siano parte di un contesto normativo socialmente riconosciuto. Questo perché è il modo in cui agiamo e interagiamo con gli altri in una rete di pratiche sociali, culturali e linguistiche che permette ai nostri stati mentali di acquisire e mostrare un'intenzionalità.

Questo significa che l'intenzionalità non può essere concepita come un fenomeno puramente mentale o individuale, ma richiede sempre una dimensione intersoggettiva e pubblica: l'adesione o il rifiuto delle norme condivise implica un processo di riconoscimento reciproco, attraverso il quale gli agenti conferiscono significato alle proprie azioni e a quelle degli altri. In questo senso, l'originarietà dell'intenzionalità non risiede tanto nella singolarità delle esperienze individuali, quanto nella capacità collettiva di creare e mantenere un ordine normativo che renda possibile l'interpretazione degli stati mentali e delle azioni come intenzionali. Il concetto di *deontic scorekeeping* riflette precisamente questa natura pubblica e relazionale dell'intenzionalità: non si tratta solamente di tenere traccia delle promesse o degli impegni degli altri, ma anche di essere continuamente coinvolti in un processo di negoziazione delle norme e dei significati all'interno di una comunità. Ciò implica che l'intenzionalità si evolve in parallelo all'evoluzione delle pratiche e dei contesti sociali, poiché le norme stesse sono soggette a reinterpretazione e adattamento. In questo quadro, l'intenzionalità individuale non è mai fissa, ma si modella costantemente in funzione delle pratiche sociali e linguistiche con cui interagisce.

Il modo in cui è possibile per l'agente umano acquisire questo tipo di intenzionalità non tramite contenuti semantici cablati nel cervello ma come effetto dell'inserimento in un orizzonte di senso condiviso è più chiaro se ci riferiamo a un altro esempio: la gestualità. Nella lettura del paradigma neo-pragmatista, la gestualità ha un significato socialmente istituito e uno stato normativo, oltre a essere uno degli elementi di cui teniamo traccia quando ci troviamo in presenza di qualcuno: «[...]

tracking the gesture is sufficient to grasp the intentionality in the other's behavior» (Gallagher 2017: 73).

Vedremo in seguito (§4.3) come alcuni dei presupposti espressi da questo approccio possano costituire un valido punto di partenza per la costruzione di un'intenzionalità enattivista, come già proposto in altre sedi (Gallagher 2017; Gallagher & Miyahara 2012; Miyahara 2011), mostrando allo stesso tempo come l'ibridazione di questo approccio con quello neo-comportamentista potrebbe fornire ulteriori possibilità di comprensione dei processi intenzionali e garantire la tenuta del concetto di intenzionalità anche quando le situazioni interazionali sono più sfumate di quelle ordinarie, come quelle tra umani e robot.

Per ora, quello che è necessario sottolineare è che questi tre approcci costituiscono, ancora oggi, le posizioni maggioritarie nel panorama filosofico quando si parla di intenzionalità. L'approccio computazionalista (neo-cartesiano), quello basato sull'attribuzione (neo-comportamentista) e quello pragmatista-fenomenologico (neo-pragmatista) veicolano differenti concettualizzazioni del fenomeno dell'intenzionalità, ma non necessariamente queste devono essere viste come mutualmente esclusive o nettamente differenti.

#### *4.1.1.4 Differenze e compatibilità*

Nell'ultima parte del suo articolo, Haugeland (1990) fornisce un breve commento sulle possibilità di tenere insieme alcuni o tutti degli approcci da lui descritti. La sua risposta è, a primo impatto, decisamente sfavorevole a qualsivoglia integrazione, «insofar as each position puts forward a *different* fundamental pattern [...] as *the* pattern upon which original intentionality depends» (Haugeland 1990: 417).

In effetti, le differenze sono decise:

- i neo-cartesiani, a differenza degli altri approcci, credono che l'intenzionalità originale sia solamente interna, proprietà esclusiva della mente confinata nel cervello dell'individuo e riducibile in potenza a contenuti interni formalmente determinati.
- i neo-comportamentisti si distinguono nel comprendere l'intenzionalità come una relazione tra agente e le azioni che realizza: attribuire intenzionalità non ha niente a che fare con la costituzione interna del sistema oggetto dell'attribuzione, e le ascrizioni non necessariamente devono essere espresse a livello linguistico per essere efficaci. Sono i comportamenti, invece, ad essere considerati fondamentali: «Acting competently in the environment takes the place of saying [...] reasonable things about it as the basic mark of the systems with intentionality» (Haugeland 1990: 414).

- I neo-pragmatisti differiscono dagli altri dal momento che pensano all'intenzionalità originale essenzialmente come una istituzione sociale, parte di una forma di vita (Haugeland 1990: 408) realizzata, mantenuta e validata dalle norme, realizzate e virtuali, della comunità di riferimento e dalle pratiche messe in atto dall'individuo nei contesti interazionali.

D'altra parte, ognuno di questi approcci mostra alcune importanti debolezze se vuole essere adottato come unico ed esaustivo approccio all'intenzionalità:

- I problemi del neo-cartesianesimo sono comuni a quelli riscontrati per l'approccio computazionalista e funzionalista delle scienze cognitive e dell'intelligenza artificiale (§2.2.1) e riguardano: i) il solipsismo derivante dall'assunto che la mente sia nascosta all'interno dell'individuo e che ii) la sua struttura sia inerentemente significativa in virtù della natura rappresentazionale della cognizione. Inoltre, pensare all'intenzionalità come proprietà che si manifesta solamente in corrispondenza di una adeguata capacità computazionale del sistema implica una concezione riduzionistica della mente, secondo cui l'intenzionalità sarebbe il risultato esclusivo di operazioni computazionali su simboli o rappresentazioni interne. Tale impostazione tende a trascurare il ruolo costitutivo del contesto ambientale e sociale nella formazione del significato (§2.2.2), riducendo l'intenzionalità a proprietà degli stati mentali che sopravviene su processi formali interni al sistema cognitivo. Gli altri due approcci contrastano la versione internalista proposta in questo paradigma aprendo a una visione che ingloba la pragmatica dell'azione e la componente sociale delle pratiche.
- L'approccio neo-comportamentista, invece, mostra una debolezza se vuole riuscire a rendere conto dell'intenzionalità originaria: se ogni agente intenzionale acquisisce la sua intenzionalità da un'attribuzione che proviene da un altro agente intenzionale, e ciò che è sufficiente e necessario per considerare l'agente tale è l'intenzionalità del comportamento esibito, dove possiamo rintracciare la forma originaria di questa intenzionalità? A meno di non ricorrere alla spiegazione per cui siamo noi stessi che ci attribuiamo un'intenzionalità – argomento che inficerebbe la definizione primaria di attribuzione da parte di altri agenti –, ciò a cui si va incontro è un processo di regresso all'infinito in cui non è possibile rintracciare l'origine dell'intenzionalità né come caratteristica della mente (Haugeland 1990), né come elemento proprio della strategia intenzionale (Adams & Aizawa 2001). Inoltre, come notato da Gallagher (2017: 69), «the idea that an agent might not have intentionality on her own seemingly undermines the idea that the external observer's attribution has its own intentionality». Il ricorso alla *folk psychology* per spiegare come le disposizioni acquistano significato e vengono utilizzate per interpretare il comportamento dell'agente attribuendogli intenzionalità risolve in



parte il problema, offrendo un quadro interpretativo intersoggettivo dell'attribuzione di stati mentali che si avvicina a quello neo-pragmatista, ma non fornisce una spiegazione soddisfacente dell'origine dell'intenzionalità, che resta comunque derivata dalle strutture interpretative dell'agente osservatore. L'attribuzione di intenzionalità basata sull'osservazione del comportamento presuppone, infatti, una forma di intenzionalità preesistente, incorrendo di nuovo in un regresso infinito. Questo circolo vizioso lascia irrisolta la questione relativa all'origine dell'intenzionalità, che non può essere interamente spiegata come proprietà derivata da pratiche di attribuzione esterna: l'approccio neo-comportamentista sembra risultare insufficiente nel fondare l'intenzionalità senza presupporre la stessa nozione che cerca di spiegare.

- Il principale elemento problematico dell'approccio neo-pragmatista è costituito dal fatto che, in questa ottica, solamente le pratiche e le azioni di individui inseriti all'interno di una rete intersoggettiva di norme e istituzioni possono essere descritte in termini di contenuti intenzionali. Questa visione lascia scoperti numerosi fenomeni in cui sembra che l'intenzionalità sia attribuita o ascritta ad animali non-umani senza capacità sociali (ad esempio, alcuni tipi di insetti), oggetti che mostrano comportamenti complessi (come quelli in cui viene implementata un'intelligenza artificiale), o anche solamente bambini che non sono ancora entrati pienamente a far parte del tessuto sociale e culturale: in altre parole, «it leaves no room whatever for original intentionality in any animals, (asocial) robots, or even isolated (unsocialized) human beings» (Haugeland 1990: 414).

Per cercare di ovviare ai singoli problemi posti ad ognuno degli approcci, alcune ibridazioni sono già state tentate, soprattutto in relazione alla necessità di prendere in considerazione non solo il funzionamento della mente del singolo, ma il suo funzionamento nel processo di comprensione di altre menti. Vedremo nel corso di questo capitolo come l'integrazione delle idee della corrente neo-cartesiana e di quella neo-comportamentista sia stata posta alla base dell'approccio più diffuso per indagare il ruolo dell'intenzionalità nell'interazione tra umani e robot, spesso contribuendo a strutturare un impianto sperimentale basato su alcuni assunti di fondo (§4.2-4.3). Per ora, resta da analizzare il modo in cui l'intenzionalità è stata considerata nel dominio degli studi sulla cognizione sociale, arrivando a costituire un elemento fondamentale da indagare per comprendere come si dia la cognizione in interazione con l'altro.

#### 4.1.2 *Intenzionalità e cognizione sociale: l'approccio classico basato sull'attribuzione*

Una comprensiva visione del ruolo dell'intenzionalità nella cognizione sociale è offerta da Malle e colleghi (2001):

Intentionality is a foundation for social cognition in several ways. For one, [i] the concept of intentionality unlocks a central part of the folk ontology of mind, because intentionality's constituent components represent basic mental categories, such as belief, desire, and awareness. Moreover, [ii] the concept of intentionality brings order to the perception of behavior in that it allows the perceiver to detect structure – intentions and actions – in humans' complex stream of movement. Further, [iii] the intentionality concept supports coordinated social interaction by helping people explain their own and others' behavior in terms of underlying mental causes. And [iv] intentionality plays a normative role in the social evaluation of behavior through its impact on assessments of responsibility and blame. (Malle et al. 2001: 1)<sup>131</sup>.

Se si analizza a fondo questo passo, ci si rende conto che tutti gli approcci all'intenzionalità analizzati da Haugeland sono raccolti qui sotto un unico quadro, ogni visione declinata all'occorrenza per puntare l'attenzione su diversi aspetti resi intelligibili da una specifica concezione di intenzionalità. Infatti, il vantaggio descritto al punto [i] sottende una nozione di intenzionalità legata ai principi esposti dall'approccio neo-cartesiano, supportata da una visione della cognizione che opera grazie a stati mentali interni alla base, i quali ne costituiscono la struttura fondazionale; i punti [ii-iii] si riferiscono, d'altra parte, alla nozione di intenzionalità per come sviluppata nell'approccio neo-comportamentista, facendo riferimento alla struttura di percezioni e azioni per rintracciare i principi del comportamento [ii] e appellandosi alla costruzione di strategie mentalistiche per predire e spiegare il comportamento dell'altro in termini di credenze e desideri [iii]; infine, il punto [iv] richiama una dimensione sociale e intersoggettiva dell'intenzionalità, che manifesta la sua efficacia soprattutto all'interno della dimensione normativa delle pratiche condivise e del giudizio sulle responsabilità dell'individuo nel contesto sociale di riferimento. La nostra analisi del ruolo dell'intenzionalità all'interno degli studi della cognizione sociale, quindi, parte dall'osservazione che i diversi approcci elencati nei paragrafi precedenti non si danno, all'interno di questo campo di studi, come definizioni mutualmente esclusive, quanto invece come alleati solidali, ognuno capace di coprire porzioni lasciate scoperte dall'altro.

Una prima differenza da operare tra gli approcci riguarda la prospettiva internalista o externalista al fenomeno, che possiamo già scorgere dalle definizioni generali di ognuno. In particolare, per l'approccio neo-comportamentista e per quello neo-pragmatista, la dimensione sociale e intersoggettiva del processo di comprensione dell'altro è già considerata come elemento costitutivo dei termini dell'intenzionalità. Da una parte, i neo-comportamentisti assumono come necessario il

---

<sup>131</sup> La numerazione ordinale dei punti non è presente nel passo originale; è stata inserita da chi scrive per favorire la fluidità dell'argomentazione ed evitare ripetizioni.

confronto di sistemi e agenti affinché l'intenzionalità possa essere attribuita, in funzione della necessità di comprendere il comportamento dell'altro ed elaborare strategie di predizione; dall'altra, i neo-pragmatisti pongono ancora di più l'accento sulla dimensione sociale, di cui quella intersoggettiva non è che un caso speciale, in cui l'intenzionalità si genera a partire dall'occupazione di una posizione socialmente data da parte dell'agente.

Diverso è il discorso per quanto riguarda l'approccio neo-cartesiano: in questo caso, la dimensione solipsistica e internalista della mente e dell'intenzionalità prescinde in termini assoluti dal rapporto intersoggettivo e sociale. Anzi, l'incontro di più individui aventi una mente e, di conseguenza, stati intenzionali che dirigono il comportamento di ognuno, si configura come nient'altro che un assemblaggio di unità discrete che non vengono influenzate dalla presenza dell'altro, la cui comprensione reciproca passa per l'inferenza degli stati mentali dell'altro. Non si farà fatica a riconoscere qui il meccanismo di cognizione sociale che abbiamo individuato come tipico degli approcci delle scienze cognitive classiche, ovvero la ToM. Infatti, l'approccio neo-cartesiano all'intenzionalità si è tradotto, negli studi della cognizione sociale, proprio nella formulazione della teoria della mente come principale impostazione concettuale attraverso cui spiegare i meccanismi di comprensione che avvengono quando la mente non è più da sola, ma si trova a doverne spiegare un'altra. Avendo già fornito una definizione dell'approccio (§3.2), qui ci preme chiarire quale sia il ruolo dell'intenzionalità all'interno della ToM e come questo si coniughi o differisca dalle altre formulazioni.

Nell'ambito della ToM, l'intenzionalità riveste una funzione essenziale poiché costituisce il fulcro del processo inferenziale mediante il quale un soggetto attribuisce stati mentali a un altro agente. In linea con la tradizione di matrice cartesiana, la ToM interpreta la comprensione dell'altro come un'attività cognitiva che consiste nell'inferire credenze, desideri e intenzioni sottostanti al – e causa del – comportamento osservabile. Tale approccio presuppone una netta distinzione tra la mente e il contesto sociale, considerando l'intenzionalità come una proprietà intrinseca di ciascun individuo, accessibile attraverso logiche basate su modelli interni, escludendo quindi una dimensione intersoggettiva e relazionale dell'intenzionalità. Possiamo caratterizzare il processo attraverso cui l'intenzionalità viene individuata nell'altro secondo questo approccio come un *processo deduttivo*: a partire dall'osservazione delle azioni dell'agente di cui ci interessa comprenderne le ragioni, si applica la logica che ci permette di determinare le cause, ovvero gli stati mentali, che generano l'azione osservata. In altre parole, dal momento che i) gli stati mentali sono visti come basi causali che guidano l'azione e ii) la comprensione delle azioni dell'altro implica l'inferenza di stati mentali, allora si può predire il comportamento dell'agente come se fosse una conseguenza logica dei suoi stati mentali.

Anche in questo processo è fondamentale il ruolo della *folk psychology*, che permette all'individuo di correlare le azioni agli stati mentali e fornire il linguaggio mentalistico attraverso cui pensare (linguaggio della mente) e parlare di credenze, desideri e intenzioni dell'agente al fine di spiegare e interpretare il suo comportamento. Come sottolineato in precedenza, l'opera di Baron-Cohen (1995), insieme a quella di Leslie (1987), rappresenta un punto di riferimento cruciale per questa prospettiva, delineando l'intenzionalità come una componente inferenziale del funzionamento cognitivo, radicata nella capacità di decodifica mentale.

La distanza più significativa da questo approccio è sicuramente quella interposta dall'approccio neo-pragmatista, in cui le pratiche intersoggettive incardinate in un sistema normativo di tipo sociale ribaltano la prospettiva, posizionandosi all'estremo opposto: l'intenzionalità è qui il risultato di pratiche linguistiche e interattive che costruiscono e negoziano costantemente il significato degli atti intenzionali e la comprensione dell'altro non si riduce alla capacità di inferire stati mentali privati, bensì alla competenza nell'interpretare ruoli sociali e normative condivise.

Rispecchiando la progressione descritta da Haugeland, esattamente a metà tra queste due posizioni si situa l'approccio neo-comportamentista, in cui abbiamo visto che la dinamica di interazione tra agenti è condizionata dall'attribuzione di intenzionalità (o dall'adozione di altre forme di atteggiamento razionalistico) che uno dei sistemi esercita sull'altro. Si può affermare che questa posizione si situi esattamente al crocevia tra gli altri due approcci proprio perché, con entrambi, il programma neo-comportamentista manifesta sia tratti in comune, sia elementi di divergenza.

- Da una parte, infatti, l'ottica neo-comportamentista sulla cognizione sociale e quella neo-pragmatista condividono una nozione di intenzionalità externalista, ovvero attribuita all'individuo dall'esterno, che sia questa derivante dalle capacità di elaborazione strategica del sistema osservatore o dalle istituzioni e dalle norme più o meno esplicitamente codificate all'interno di un ambiente sociale, riservando anche un ruolo significativo alla pragmatica dell'azione e alle capacità dell'agente di mettere in pratica le norme provenienti dal contesto culturale – attraverso la *folk psychology* per l'atteggiamento intenzionale, o nel rispetto del *deontic skorekeeping* nell'ottica pragmatista; allo stesso tempo, le prospettive differiscono considerevolmente nella considerazione degli stati mentali nel processo di comprensione dell'altro e nell'oggetto da considerare per indagare il fenomeno della cognizione sociale – per i primi il comportamento, per i secondi la comunicazione linguistica, l'adesione alle norme sociali e la pragmatica dell'azione.
- Dall'altra parte, l'approccio neo-cartesiano condivide con quello neo-comportamentista l'idea che l'intenzionalità sia cruciale per comprendere e spiegare il comportamento umano, e che

questa vada correlata agli stati mentali interni del sistema cognitivo. Gli elementi di disaccordo riguardano principalmente le modalità attraverso cui comprendere tale correlazione:

- per i neo-cartesiani il legame tra stati mentali e azioni è causale, secondo una direzionalità che va dai primi alle seconde: il processo inferenziale presuppone che il comportamento del sistema non sia intenzionale di per sé, ma deve essere interpretato tramite la mediazione di un processo cognitivo interno.
- per i neo-comportamentisti, l'intenzionalità è direttamente osservabile nel comportamento ed è a partire da tale osservazione che è possibile costruire una strategia che attribuisce intenzionalità anche agli stati mentali dell'agente. Il comportamento viene compreso come direttamente intenzionale grazie alla significatività individuata nelle risposte osservabili. Il processo inferenziale, al contrario di quello neo-cartesiano, è *induttivo*: a partire dall'osservazione di comportamenti specifici si costruisce una generalizzazione empirica degli stati mentali. Ciò, non implica alcun assunto ontologico, dal momento che gli stati mentali non devono essere necessariamente né giusti, né interni all'agente, ma solamente adeguati a supportare l'atteggiamento adottato: l'intenzionalità è quindi da considerare come il prodotto di regolarità comportamentali visibili e pattern d'azione.

Ora, va notato che l'approccio neo-pragmatista, salvo alcune eccezioni che vedremo in seguito (§4.2.3), è quello che ha avuto meno fortuna all'interno delle scienze cognitive, complice probabilmente la popolarità guadagnata dall'approccio neo-cartesiano e l'importante divergenza manifestata tra le due visioni, la quale potrebbe aver contribuito a far scivolare in secondo piano le considerazioni sulla struttura sociale e normativa dell'intenzionalità anche nella cognizione sociale.

È invece la seconda delle convergenze possibili che è stata sviluppata, fino ad assumere il ruolo di teoria maggiore dell'intenzionalità all'interno del panorama degli studi sulla cognizione sociale, attraverso il minimo comun denominatore della *folk psychology*. Come si è visto, in entrambi gli approcci la psicologia del senso comune risulta fondamentale o per strutturare la base logico-linguistica tramite cui ragionare sugli stati mentali degli altri (neo-cartesiani), o per fornire gli elementi attraverso cui l'individuo forma le proprie credenze e le mette in relazione a creare una strategia di comprensione (neo-comportamentisti). È sotto la congiunzione resa disponibile dalla *folk psychology* che è possibile rileggere i punti [ii-iii] – «[ii] the concept of intentionality brings order to the perception of behavior in that it allows the perceiver to detect structure – intentions and actions – in humans' complex stream of movement»; «[iii] the intentionality concept supports coordinated social interaction by helping people explain their own and others' behavior in terms of underlying

mental causes» – annoverati da Malle e colleghi (2001: 1) come comuni a entrambe le visioni riguardo l'intenzionalità nella cognizione sociale: il processo di attribuzione di intenzionalità al sistema si mischia con gli assunti filosofici dell'approccio neo-cartesiano sul dualismo e sull'interiorità della mente. In sostanza, il processo di comprensione dell'altro prende la forma dell'incontro di due menti confinate all'interno dell'individuo che per comprendersi si attribuiscono intenzionalità a vicenda, non più attraverso inferenze induttive che vanno dal comportamento ai potenziali stati mentali, ma come una deduzione che cerca di risalire agli stati mentali intenzionali che hanno portato al comportamento osservato.

Mettendo da parte le divergenze<sup>132</sup>, quindi, la ToM è diventata il paradigma comune attraverso cui descrivere il processo di *attribuzione di intenzioni, credenze e desideri agli altri, ragionando circa il legame causale tra stati mentali e azioni esibite*. È soprattutto con l'intento di affrontare il problema della cognizione sociale nelle scienze cognitive, infatti, che si delinea l'alleanza tra i proponenti di un approccio inferenziale alla cognizione e i teorici dell'attribuzione dell'intenzionalità: così, i neo-cartesiani guadagnano la possibilità di declinare in ottica intersoggettiva le attività delle mente, che comunque resta confinata all'interno dell'individuo; i neo-comportamentisti trovano una soluzione al regresso all'infinito dell'intenzionalità originale, che trova spazio all'interno della mente dell'individuo ma che acquista significato solamente se attribuita da altri sistemi<sup>133</sup>.

In sintesi, l'attribuzione di intenzionalità tramite meccanismi di ToM è diventata il modello più forte all'interno delle scienze cognitive per spiegare il processo di comprensione che avviene nella cognizione sociale, sviluppandosi in diverse direzioni e coprendo domini che vanno dalla classica psicologia sociale (Malle & Knobe 1997; Malle 2011) all'etologia (Premack & Woodroof 1978; Povinelli 1994; Krupenye & Call 2019), dalla psicologia dello sviluppo (Rosati et al. 2001; Wellman 2010; Tomasello 2018) alla filosofia della mente (Malle et al. 2001, Eds.), fino agli studi sulle psicopatologie come l'autismo (Baron-Cohen et al. 1985) e la schizofrenia (Frith 1994).

---

<sup>132</sup> I modi attraverso cui è possibile coniugare i due approcci in modo da far risultare le differenze meno rilevanti è analizzato da Malle (2005). Malle (2005: 11) distingue quattro modalità di integrazione, ma tutti pongono il meccanismo degli stati mentali sulle azioni e i comportamenti dell'agente come base del processo di attribuzione di intenzionalità: «In sum, the concept of intentionality spans four domains of explanation and their corresponding modes. When intentionality is not ascribed, people offer cause explanations. When intentionality is ascribed, people offer either reason explanations, causal history of reason explanations, or enabling factor explanations».

<sup>133</sup> Haugeland fa riferimento ad alcune posizioni di Dennett (1987) per motivare una possibile ibridazione tra approccio neo-cartesiano e neo-comportamentista, come la possibilità di integrare l'ipotesi del linguaggio del pensiero all'interno della teoria delle *stances*. Tuttavia, lo stesso Haugeland tiene a sottolineare che il modo in cui i due approcci potrebbero trovare un compromesso non rende trattabile l'assunto di Dennett che non ci sia corrispondenza tra l'assunzione dell'atteggiamento intenzionale e la struttura fisica e materiale dell'implementazione nell'agente: «[...] even if a neo-behaviorist were to allow internal computational tokens at the implementation level, these needn't be identical with *either* the system's "ordinary" cognitive states, intervening between perceptions and actions, *or* its "verbal cognitions" (if any) formulated explicitly in its linguistic module. The internal tokens might very well have some sort of intentionality or representational content, but it would be (in the present compromise, anyway) metaphysically derivative» (Haugeland 1990: 425, nota 38). Torneremo in seguito sulla questione dell'incompatibilità tra gli approcci in oggetto (§4.2.3). Vedremo a breve come Dennett (1991, 1995) avesse esplicitamente rifiutato l'ipotesi di un LoT interno alla mente.

Il dominio degli studi sull'interazione tra umani e robot non ha fatto eccezione; anzi, negli ultimi anni la HRI è stato uno dei principali campi disciplinari all'interno del quale gli studiosi di intenzionalità hanno messo alla prova le proprie teorie e hanno elaborato metodologie di indagine e verifica dei meccanismi di attribuzione e inferenza. Replicando il successo riscosso dalla ToM per spiegare i meccanismi di comprensione del robot come agente sociale, anche nel caso dello studio dell'attribuzione di intenzionalità nei confronti del robot il modo attraverso cui viene inquadrata la questione passa per gli strumenti teorici offerti dalla teoria della mente.

A seguito della nostra analisi sul concetto di intenzionalità nelle scienze cognitive, però, abbiamo rintracciato divergenze significative fra gli approcci, soprattutto in merito alla questione dell'intenzionalità originaria e derivata, in relazione al ruolo e alla natura degli stati mentali in questo processo e agli strumenti normativi che possono essere forniti per decifrare l'adeguatezza della nostra comprensione dell'altro. Nel caso che ci interessa, ovvero quando l'altro non è un essere umano ma un robot sociale, i meccanismi teorizzati nell'incontro tra idee neo-cartesiane e neo-comportamentiste, elaborati allo scopo di spiegare il processo di mutua comprensione attraverso l'impiego del concetto di intenzionalità, reggono alla sfida degli agenti non umani?

Questo non è l'unico interrogativo che emerge se trasferiamo le questioni che abbiamo appena trattato al contesto della HRI: esistono differenze significative nei processi di attribuzione di intenzionalità quando ci rapportiamo a un agente umano rispetto a un agente artificiale? Possiamo parlare di un'intenzionalità *autentica*, o si tratta di un'intenzionalità derivata, legata a schemi di inganno percettivo che inducono a proiettare intenzionalità in sistemi che non ne sono intrinsecamente dotati? E se attribuiamo intenzionalità a un robot, in che misura tale attribuzione è epistemicamente giustificata e quali conseguenze ha per il nostro modo di spiegare i suoi comportamenti? Non crediamo che domande come queste siano interessanti solamente dal punto di vista della speculazione filosofica: esse si collocano al centro del dibattito sull'HRI e hanno profonde ricadute sul piano della spiegazione intenzionale. La percezione di un robot come agente intenzionale può influenzare non solo la natura dell'interazione, ma anche la maniera in cui interpretiamo e giustifichiamo i suoi comportamenti, potenzialmente rimodellando le nostre aspettative e la nostra possibile cooperazione con esso.

Per ora, rispettando i termini del dibattito per come impostato classicamente dai modelli di cognizione sociale adottati e dalle concettualizzazioni degli strumenti epistemicici in gioco, l'alternativa è tra due posizioni poste agli estremi: se, infatti, l'attribuzione di intenzionalità è il risultato di un inganno percettivo, ciò pone seri interrogativi sulle modalità con cui i sistemi robotici possono essere integrati all'interno delle dinamiche sociali umane, e sulle implicazioni epistemologiche di tali attribuzioni. Al contrario, se tale intenzionalità può essere vista come il

risultato della rivelazione di un pattern significativo in cui l'attribuzione è una possibilità confermata dagli eventi, allora questa può avvenire anche in assenza di correlati ontologici localizzabili nella struttura cognitiva dell'agente. Anticipiamo che il nostro obiettivo consiste nel mostrare che, se l'intenzionalità è effettivamente rintracciata nell'interazione tra robot e umani, si può aprire la strada a una riconsiderazione radicale delle modalità attraverso cui la nostra cognizione sociale si adatta alla presenza di agenti non umani.

## **4.2 La mente che non c'è: attribuzione di intenzionalità negli HRI**

### *4.2.1 L'importanza di essere "agente intenzionale"*

Riuscire a comprendere se e come gli esseri umani attribuiscono intenzionalità ai robot è una questione che acquista notevole rilevanza non solamente sul piano teorico, per la messa alla prova delle teorie su intenzionalità e interazione sociale, ma costituisce anche uno snodo fondamentale nel contesto di concreta inclusione dei robot nelle nostre pratiche interazionali quotidiane e apre a possibilità di intervento nel processo di miglioramento e ottimizzazione della presenza di questi agenti artificiali nel nostro tessuto culturale. Considerare il robot come *agente intenzionale*, infatti, significherebbe poter fare affidamento su un apparato di capacità e competenze che il robot riuscirebbe a mettere in pratica (Perez-Osorio & Wykowska 2019). Infatti, se i robot fossero capaci di suscitare la sensazione di essere effettivamente capaci di assurgere ai compiti sociali e interazionali suggeriti dalle loro capacità mimetiche e sintetiche, significherebbe che possono essere considerati agenti alla pari degli esseri umani e potrebbero tranquillamente prendere il posto di qualsiasi altro agente con capacità sociale che ci è così familiare: diventare, in sostanza, *sostituti* (Dumouchel & Damiano 2017, cfr. §3.2).

In particolare, come notato altrove (Dautenhahn 2007, Fong et al. 2003; Perez-Osorio & Wykowska 2019; Thellman et al. 2017; Marchesi & Wykowska 2023), ciò aprirebbe una serie di possibilità per il robot, che altrimenti resterebbero escluse. In primo luogo, considerare il robot come un agente intenzionale equivale a supportarne la comprensione delle azioni e a prevederne gli sviluppi: ciò faciliterebbe la capacità di *attunement* tra gli agenti coinvolti, dato che i robot costituirebbero solamente un altro esempio di agente sociale. Di conseguenza, questo comporterebbe un forte aumento delle capacità di cooperazione tra robot ed esseri umani (Yamazaki et al. 2010; Carter et al. 2011), i quali non farebbero fatica a coordinarsi alle azioni del robot. Inoltre, i robot potrebbero sfruttare canali di comunicazione più impliciti di quelli che solitamente adottano, come la comunicazione verbale e l'esasperazione delle espressioni facciali: la comunicazione non verbale, ad esempio, potrebbe essere implementata così da rendere l'interazione più naturale. Infine, si potrebbe pensare che considerare il robot come agente intenzionale genuino comporti dei vantaggi per la sua



accettazione sociale, oltre che per mettere in atto processi affettivi ed empatici che ne potenzierebbero le capacità di intervento e per istaurare rapporti di fiducia (Kiesler et al. 2008; Carter et al. 2011).

Comprendere, quindi, *se* davvero gli esseri umani attribuiscono intenzionalità ai robot e *come* questo avvenga in termini di processi cognitivi e capacità interazionali è diventata una delle prerogative degli studi sulla human-robot interaction. Soprattutto negli ultimi anni, in cui la presenza di robot si è allargata non solamente in contesti di interazione diretta con l'essere umano, ma anche a forme meno esplicite e antropomorfe (come, ad esempio, nel caso delle auto a guida autonoma; Petrovych et al. 2018; Thellman & Ziemke 2020; Ziemke & Bischoff 2023), la necessità di comprendere i meccanismi del processo di attribuzione di intenzionalità si è rivelato cruciale per direzionare la ricerca e la costruzione di agenti artificiali in un modo che sia il più conforme possibile alle nostre tendenze di attribuzione (Banks 2020; Chaminade et al., 2012; Cucciniello et al. 2023; Krach et al., 2008; de Graaf and Malle, 2017, 2018, 2019; Marchesi et al., 2019, 2021a; Terada et al., 2007; Thellman, 2021, 2021; Thellman and Ziemke, 2019, 2020; Ziemke, 2023; Thellman et al. 2022; Thellman et al. 2017; Ziemke et al. 2015; Wykowska 2021).

Se vista nei termini che abbiamo delineato poco sopra, l'attribuzione di intenzionalità o di stati mentali a sistemi artificiali, oltre ad avere conseguenze ideali sulle possibilità di inclusione del robot nella comunità umana e all'aumento delle possibilità di comprensione, si configura come la possibilità per l'essere umano di predire, spiegare e interpretare il comportamento dell'agente in oggetto in termini di credenze e desideri, di percezioni e intenzioni, e non solamente in riferimento a spiegazioni fisiche o meccanicistiche relative al suo funzionamento (cfr. Imamura et al. 2015; Perez-Osorio & Wykowska 2019). Proseguendo questo ragionamento, ciò significherebbe che il robot è visto come un essere umano, guidato allo stesso modo dai propri stati mentali, e di conseguenza il suo comportamento diventa interpretabile non più solamente in funzione della sua architettura o struttura materiale, ma in relazione a ciò che lo anima dall'interno, alle disposizioni mentali che rimangono – secondo la ToM – invisibili all'interlocutore.

Ora, all'interno degli studi sulla HRI, il fenomeno dell'attribuzione dell'intenzionalità è stato preso in considerazione attraverso diverse etichette, come *mental state attribution/ascription* (Appel et al. 2020; Abubshait & Wykowska 2020; Gray & Wegner 2012; Mutlu et al. 2009; Terada et al. 2007), *anthropomorphism* (Epley et al. 2007; Zlotowski et al. 2017, 2018; anche detto *psychological anthropomorphism*, cfr. Eyssel & Reich 2013; Kamide et al. 2013), *mind perception* (Desideri et al. 2021; Gray et al. 2007; Wiese et al. 2021), *theory of mind* (Scassellati 2002), *mentalizing* (van Overwalle & Baetens 2009; Abubshait & Wykowska 2020), *intentional stance* (Marchesi et al. 2019, 2021a; Chaminade et al. 2012, 2018; Wiese et al. 2019; Wykowska et al. 2014; Perez-Osorio & Wykowska 2020, 2019), *folk psychology* (de Graaf & Malle 2019; Fiala et al. 2014; Thellman &

Ziemke 2020; Larghi & Datteri 2024) e *mindreading* (Lee et al. 2021; Ito et al. 2004; Terada & Yamada 2017). Come notato da Thellman e colleghi (2022), i termini però non sembrano denotare concetti diversi, quanto invece illuminare aspetti complementari di uno stesso fenomeno, ovvero quello dell'attribuzione/ascrizione di stati mentali ad agenti robotici<sup>134</sup> concepito come «the cognitive capacity to reflect upon one's own and other persons' mental states such as beliefs, desiders, feelings and intentions» (Brüne et al. 2007: 151).

La mancanza di consenso non si arresta alla scelta del nome per indentificare il fenomeno, ma si allarga anche i) all'incertezza sul meccanismo che lo costituisce e ii) alle modalità di verifica della presenza e validazione del meccanismo stesso. Per quanto riguarda i), diverse ricerche (Duffy 2003; Hortensius & Cross 2018; Zlotowski et al. 2015) forniscono evidenze che gli esseri umani descrivano correntemente il comportamento del robot in termini mentalistici, attribuendo dunque – in teoria – stati mentali all'agente artificiale. Queste ricerche, tuttavia, sono in contrasto con un'altra serie di studi che evidenziano come, se chiesto esplicitamente, gli esseri umani negano che i robot che hanno incontrato posseggano una mente o stati mentali che ne possano causare il comportamento (Banks 2020; Fussell et al. 2008). Inoltre, altri esperimenti suggeriscono che non sempre il modo in cui un soggetto descrive le azioni di un robot ha corrispondenza nell'atteggiamento adottato verso di lui. Ad esempio, Bartneck e colleghi (2007; Bartneck & Hu 2008) hanno chiesto ai soggetti umani di colpire il robot con un martello o spegnerlo: nella maggior parte dei casi, le persone si sono rifiutate di compiere l'azione richiesta o hanno mostrato notevole difficoltà ed esitazione, ma allo stesso tempo, alla domanda se credessero che il robot potesse sentire dolore, morire o possedere una mente, hanno risposto negativamente. In sostanza, c'è una notevole discrepanza tra il modo in cui viene impiegato il linguaggio mentalistico e le disposizioni nei confronti dell'agente robotico, fattore che rende complessa l'indagine dei meccanismi cognitivi che sottostanno alla attribuzione di stati mentali ai robot.

Le discordanze non si appianano nemmeno nei riguardi di ii). Infatti, l'apparato sperimentale con cui si vaglia la possibilità che gli esseri umani attribuiscano stati mentali ai robot e si cerca di comprenderne le modalità di sviluppo del meccanismo cognitivo si configurano principalmente in tre strategie. La prima, senza dubbio la più diffusa, consiste nel somministrare questionari in forma verbale agli individui che, nella maggior parte dei casi, devono esprimere il proprio livello di accordo all'affermazione (*Likert scale*, Marchesi et al. 2019; Imamura et al. 2015; Cross et al. 2019; o scala semantica differenziale, cfr. Spatola & Wudarczyk 2021; Schellen & Wykowska 2019), oppure fornire delle descrizioni delle azioni compiute dal robot (Short et al. 2010; de Graaf & Malle 2019) o

---

<sup>134</sup> Accogliamo quindi il proposito di Thellman et al. (2022) di coniugare le diverse accezioni sotto l'etichetta di "attribuzione di stati mentali". Da qui in avanti, la denotazione verrà impiegata per identificare il fenomeno insieme a quella di "attribuzione di intenzionalità".

predirne gli esiti (Banks 2020; Levin et al. 2012, 2013). Questa misura è uniformemente considerata una forma *esplicita* di validazione della procedura di attribuzione di intenzionalità ai robot, secondo cui optare per una descrizione mentalistica o adottare un linguaggio che include stati mentali per descrivere, spiegare o predire le azioni del robot corrisponde all'adozione di un atteggiamento intenzionale e, di conseguenza, alla considerazione del robot come agente intenzionale.

La seconda modalità, anche questa comune ma sicuramente meno diffusa della prima<sup>135</sup>, consiste nell'impiego di misure comportamentali, anche dette *implicite*. In questi casi, l'attribuzione di stati mentali ai robot da parte di agenti umani viene validata tramite test che prevedono che l'agente umano si comporti con il robot come se fosse di fronte ad un altro essere umano, fornendo quindi evidenza dell'attribuzione di stati mentali all'interlocutore e del suo status di agente intenzionale. In particolare, esempi di misurazioni comportamentali sono quelli in cui si considera la tendenza dell'agente umano di adottare un comportamento sociale nei confronti del robot (Lemaignan et al. 2015; Straub 2016) o la disponibilità di aiutare il robot a raggiungere i propri obiettivi (Martin et al. 2020; Yamaji et al. 2010), ma anche quelli in cui si monitora la performance dell'agente umano in interazione con il robot, il cui esito è dipendente dall'attribuzione di stati mentali. Esempi di questo ultimo scenario sperimentale includono la verifica dell'adozione della prospettiva del robot da parte dell'essere umano, il quale si mette dunque nei “suoi panni” (Zhao et al. 2016), o la presenza dello sguardo anticipatorio (*anticipatory gaze*, §4.3.3-4.4.2) nel comportamento umano (Sciutti et al. 2013; Thellman & Ziemke 2020).

Infine, l'ultima modalità adottata è quella che prende in considerazione le evidenze neurobiologiche per comprendere se, alla vista di un robot sociale, si attivino gli stessi pattern neuronali in uso nelle interazioni sociali con gli esseri umani. Soprattutto attraverso l'impiego della fMRI (*functional Magnetic Resonance Imaging*, ad es. Cross et al. 2019; Chaminade et al. 2012; Gobbini et al. 2011), ma anche tramite EEG (elettroencefalografia, cfr. Bossi et al. 2020), alcuni ricercatori hanno mostrato come l'attivazione neuronale di fronte a contesti in cui è innescata la risposta empatica è simile sia per scenari che comprendono solamente esseri umani, sia solamente robot (Rosenthal-von der Pütten et al. 2014), o per misurare i livelli di ‘mentalizzazione’ del robot in contesti come quello del “dilemma del prigioniero” (Hegel et al. 2008), ma anche per mostrare come alcune aree ipotizzate essere fondamentali per la comprensione intersoggettiva – quali la corteccia temporale, e l'attivazione subcorticale nell'ipotalamo, nel talamo, nell'ippocampo, nell'amigdala e

---

<sup>135</sup> In uno studio sistematico riguardo lo stato dell'arte del fenomeno in oggetto, Thellman et al. (2022) riportano che, su un campione di 155 esperimenti, 130 utilizzano modalità verbali di verifica e valutazione del processo di attribuzione di stati mentali, utilizzando questionari o altre misure esplicite, mentre solamente 21 studi impiegano misure comportamentali e 14 misure neurologiche (cfr. Thellman et al. 2022: 14, fig.5).

nell'area subtalamica – si attivino solamente nel caso di interazione (verbale) con un altro essere umano e non con un robot (Rauchbauer et al. 2019).

Sorgono quindi problemi di non facile risoluzione riguardo alla possibilità di integrare le diverse forme di misurazione e validazione del processo di attribuzione di intenzionalità ai robot, soprattutto alla luce del fatto che le evidenze sono spesso in contraddizione tra di loro: da una parte, infatti, alcuni studi testimoniano di una tendenza diffusa a parlare dei robot in termini mentalistici – anche se influenzata da molti fattori come l'età (Di Dio et al. 2018, 2020; Levin et al. 2013), il contesto culturale di riferimento (Tan et al. 2018; Trovato & Eyssel 2017), il *gender* di appartenenza (Tan et al. 2018; Searbeck & Bartneck 2010) – mentre altri riportano una difficoltà nel validare lo stesso meccanismo attraverso misure implicite che non coinvolgano necessariamente la verbalizzazione. È facile immaginare uno scenario in cui questo possa avvenire anche a noi: per quanto io sia affezionata al mio computer, che mi consente di fare ricerca in modo agile e scrivere la mia tesi di dottorato, non sarei disposta ad ammettere che abbia un'anima, una mente o che quando lo faccio cadere senta dolore, anche se probabilmente gli chiederei se si sia fatto male. In effetti, sembra questa la tendenza più comune, quella per cui parliamo ai nostri dispositivi, ancor più se mostrano tratti mimetici e sintetici simili a quelli di un essere umano, come se fossero nostri amici, anche se non saremmo mai disposti ad ammettere che abbiano una mente (cfr. Özdem et al. 2017). Come anche riportato da Thellman, de Graaf e Ziemke (2022: 2),

Many people might say that their robot lawnmower wants to avoid colliding with trees, although they would not say it has a mind, a will, or desires. In other words, it is not uncommon to conceptualize the behavior of robots as mind-governed without necessarily believing that robots really have minds, similar to how we interpret the behavior of fictional characters, companies, and nation-states.

Gli autori di questo studio, a seguito di un'approfondita analisi sui fattori distintivi della ricerca sull'attribuzione di stati mentali ai robot sociali, identificano una serie di tratti comuni all'intero impianto di indagine, riassumibili in quattro punti (Thellman et al. 2022: 28):

- a) Nonostante la disomogeneità nell'identificazione del fenomeno, testimoniata dalla proliferazione di nomi per riferirvisi in letteratura, gli studi mostrano una comune attenzione per il ruolo degli stati mentali nel processo di attribuzione di intenzionalità, tanto da far propendere a uniformare il campo di studi sotto l'etichetta di “attribuzione di stati mentali”.
- b) La tendenza ad attribuire stati mentali ai robot viene considerata determinata da fattori come l'età e gli obiettivi degli agenti umani, da una parte, e dal comportamento e dalle sembianze del robot, dall'altra.
- c) È rintracciabile la propensione a classificare gli agenti dal più al meno intenzionale sulla base di una nozione ingenua di *embodiment* (che si potrebbe ricondurre al tipo di *embodiment fisico*

e *organismoide* elencati in §3.4.2), che si ferma alla considerazione del livello superficiale relativo alla forma dell'agente, cioè alla componente estetica del mimetismo istanziato dal robot. Per cui, la gerarchia che si delinea è quella *computer < robot < umano*, in cui la tendenza ad attribuire gli stati mentali sembra essere correlata alla presenza di un aspetto antropomorfo.

- d) Infine, gli autori notano una sostanziale incoerenza tra i risultati nella letteratura empirica, imputata all'utilizzo di metodologie eterogenee in cui le misure implicite e quelle esplicite vengono alternativamente sviluppate in solitaria, o ibridate senza il necessario sforzo teorico di sistematizzazione.

A queste considerazioni, proviamo a integrare le nostre, sulla scorta del percorso che abbiamo compiuto fino ad ora. Per quanto riguarda a), il motivo per cui è possibile raccogliere sotto l'etichetta prescelta di "attribuzione di stati mentali" il fenomeno che stiamo osservando è relativo alla comune adozione, da parte di tutti gli studi presi in oggetto, dell'approccio dell'attribuzione/ascrizione dell'intenzionalità riletto alla luce della ToM. Questo studio non fa altro che ribadire la già registrata pervasività del modello della teoria della mente negli studi sull'interazione tra umani e robot, il quale detta i propri termini nello sviluppo della maggior parte – se non nella totalità – delle indagini sul tema e fornisce i propri strumenti per declinarne lo studio e la comprensione: ovviamente, l'intenzionalità e il suo ruolo nella comprensione dei robot non hanno fatto eccezione. In merito a b), non è tanto la presenza di una serie di tratti che influenzano il processo, tutti provati essere dirimenti in determinate circostanze sia nella comparsa che nelle modalità del processo di attribuzione, ma una mancanza che vogliamo sottolineare. Infatti, tra tutti i fattori elencati, non si fa menzione dell'interazione come elemento che, anche solo potenzialmente, può influire sul fenomeno: il ruolo dell'interazione in vivo è del tutto trascurato, e questo non per negligenza degli autori dello studio. Data l'importanza del campione preso in considerazione (155 studi pubblicati su riviste internazionali di settore), la mancanza del fattore-interazione tra gli elementi considerati capaci di intervenire nel processo di formazione dell'attribuzione fa ben comprendere come la componente incarnata e situata non sia ancora affatto rilevante negli studi specialistici (come sottolineato anche in §3.2-3.3). Anche in c) ritroviamo sottotraccia la mancanza di attenzione verso un contesto situato dell'interazione, che si riverbera non solo nella considerazione limitata delle possibilità interazionali dell'agente robotico – che riflette la superficiale composizione della gerarchia che va da computer a essere umano, elaborata semplicemente sulla base delle possibilità sociali aperte dalla forma dell'*embodiment organismoide* realizzato e dalle forme estetiche dell'agente –, ma anche e soprattutto nel contesto sperimentale più comune. Come già sottolineato, nella maggior parte dei casi il contesto sperimentale

non prevede una reale interazione con il robot: non c'è possibilità di cooperare per raggiungere un obiettivo comune, non c'è contatto fisico o scambio verbale non programmato, non ci sono possibilità di improvvisazione per alcuna delle parti coinvolte né dal punto di vista conversazionale, né corporeo. Il modello della misurazione esplicita prevede nella maggioranza dei casi la visione di video e foto, corredata dalla compilazione di un questionario in cui optare per una descrizione con termini mentalistici che guida un'interpretazione compiacente nei confronti della teoria della mente e degli strumenti che mette a disposizione. Infine, al punto d) riserveremo il seguente paragrafo, in cui analizzeremo uno dei *setting* sperimentali più adottati in HRI per lo studio del fenomeno dell'attribuzione dell'intenzionalità e cercheremo di comprendere come questo provi a integrare i risultati discordanti degli esperimenti legandosi ai presupposti teorici che, almeno a prima vista, sembrano adottare.

#### 4.2.2 *La misura dell'intenzionalità*

Uno degli esempi attuali più significativi rintracciabili in letteratura circa la validazione della presenza dell'atteggiamento intenzionale verso i robot sociali e in linea con l'approccio più diffuso dell'attribuzione di stati mentali è quello sviluppato da un gruppo di ricercatori all'unità "Social Cognition in Human-Robot Interaction" (S4HRI) dell'Istituto Italiano di Tecnologia di Genova. La linea di ricerca ha esattamente lo scopo di comprendere se e come gli esseri umani attribuiscono intenzionalità e stati mentali ai robot, conducendo numerosi esperimenti e contribuendo in modo sostanziale all'aumento di evidenze e conoscenza circa il fenomeno dell'adozione dell'atteggiamento intenzionale nei confronti di agenti artificiali (Marchesi & Wykowska 2023; Perez-Osorio & Wykowska 2019, 2020; Wykowska 2024 (Ed.)).

Al fine di analizzare l'impianto sperimentale più diffuso per la comprensione del fenomeno dell'attribuzione dell'intenzionalità, il risultato più significativo da prendere in considerazione è l'*InStance Test* (IST). Come riportato e validato in diversi articoli (Spatola et al. 2021a, 2021b; Abubshait & Wykowska 2020; Marchesi 2021; Marchesi et al. 2019, 2021a, 2021b, 2021c, 2024; Ghiglini et al. 2023; Roselli et al. 2023), il test consiste in un questionario (*InStance test Questionnaire*, ISQ) in cui i soggetti umani partecipanti all'esperimento devono scegliere se optare per una descrizione meccanicistica o mentalistica del comportamento del robot che osservano. Nello specifico, lo studio elaborato da Marchesi e colleghi (2019) per la prima validazione del test espone in dettaglio l'apparato sperimentale e i presupposti teorici ed epistemici che ne guidano l'interpretazione.

Come si può notare fin dal titolo<sup>136</sup>, la proposta adotta da subito l'approccio dell'atteggiamento intenzionale per come elaborato da Dennett (1971, 1987), riportando citazioni dei testi in cui l'*Intentional stance* è descritta dall'autore e specificando la differenza con gli altri tipi di atteggiamenti, in particolare con quello progettuale (Marchesi et al. 2019: 2). Infatti, una delle affermazioni che guidano la strutturazione del questionario è relativa alla necessità per l'agente umano di adottare una delle *stance* di fronte al sistema robotico, e la relativa scelta può rendere conto di come il robot è considerato in interazione: «[...] if an empirical test is set out to examine whether one adopts the intentional stance toward a system, the constrasting conditions should be either the design stance or the physical stance» (*ibid.*). Il test, quindi, si struttura a partire dai presupposti classicamente delineati da Dennett riguardo l'adozione di uno degli atteggiamenti al fine di elaborare interpretazioni del comportamento del sistema.

Un altro elemento che viene adottato a partire dalla teoria di Dennett consiste nel non considerare necessariamente il sistema robotico come un agente in possesso di un'intenzionalità genuina. Gli autori, infatti, abbracciano la visione riguardo la non obbligatorietà che il sistema osservato possieda una intenzionalità originale, che sia quindi effettivamente un sistema intenzionale a sua volta – un *true believer*. Questo tratto della teoria è, probabilmente, quello che l'ha resa l'approccio teorico più diffuso all'interno di questo tipo di indagini, poiché consente di studiare il fenomeno senza necessariamente affermare l'esistenza di un'intenzionalità interna al sistema robotico, svincolando quindi gli studi da analisi sulle conseguenze ontologiche che la rilevazione di attribuzione di intenzionalità agli agenti artificiali comporta nell'account computazionalista:

In this context, it needs to be noted, however, that adopting the *intentional stance* toward an artifact (such as a humanoid robot) does not necessarily require that the artifact itself possess true intentionality. Adopting the *intentional stance* might be a useful or default way to explain a robot's behavior. Perhaps we do not attribute mental states to robots but we treat them *as if* they had mental states. (Marchesi et al. 2019: 3).

Inoltre, più volte gli autori dello studio rimarcano che la loro indagine è volta a comprendere: i) «whether some contexts can evoke mentalistic explanations of behavior of a humanoid robot» (ivi: 3) e ii) a sviluppare «a method for assessing whether humans would sometimes (at least in some contexts) adopt the intentional stance toward the humanoid robot iCub» (ivi: 8). Per quanto riguarda i), la presenza di una descrizione in termini mentalistici del comportamento del robot implicherebbe l'adozione da parte dell'agente umano dell'atteggiamento intenzionale, ovvero quella strategia basata su termini che si riferiscono a stati mentali come credenze, desideri e percezioni utile a interpretare e predire il comportamento del sistema osservato; d'altra parte, se l'individuo opta per una descrizione che non contiene riferimenti a stati mentali, ma solamente al funzionamento meccanico interno del

---

<sup>136</sup> “Do we adopt the intentional stance toward humanoid robots?” (Marchesi et al. 2019).

robot, allora significa che l'*intentional stance* non è stata adottata e che il robot non è visto come un agente intenzionale, ma solo come una macchina molto complicata, interpretabile – al massimo – secondo le logiche della *design stance*.

A partire da questi presupposti, dunque, si procede allo sviluppo di ii): il *setting* dell'ISQ consiste in una serie di immagini, mostrate al soggetto dell'esperimento, che ritraggono il robot umanoide iCub (Metta et al. 2010) svolgere azioni differenti in uno scenario neutrale. L'individuo si trova, quindi, davanti ad una sequenza di tre immagini che rappresenta lo svolgimento dell'azione e per ogni scenario deve compiere una scelta: ad un estremo della schermata appare una descrizione delle azioni del robot contenente termini mentali, mentre all'altro estremo l'affermazione riporta una descrizione in termini meccanici.

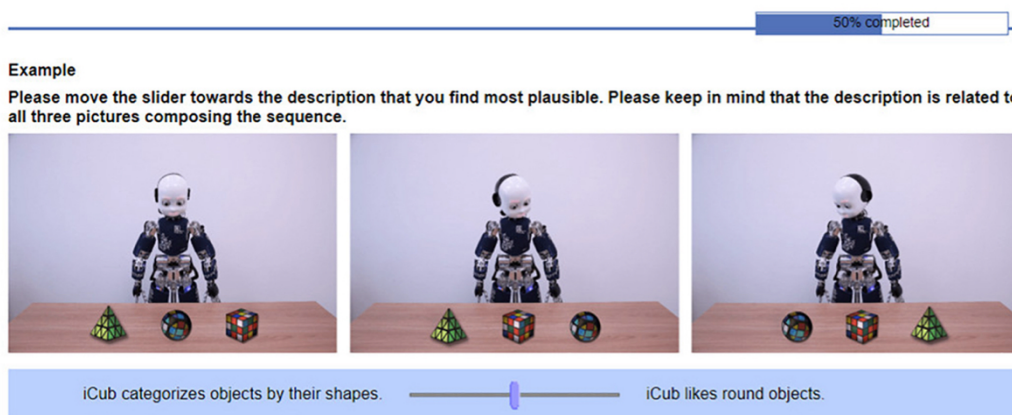


Figura 4.1. Schermata di un passaggio dell'*Instance Test Questionnaire* in inglese. (Marchesi et al. 2019: 5)

In Fig. 4.1 vediamo come queste due diverse posizioni siano esemplificate dalla opposizione semantica “*categorizes vs likes*”: mentre il primo termine si riferirebbe a un funzionamento interno guidato da meccanismi o algoritmi che dipingono il robot nei panni di una macchina artificiale intelligente, nel secondo caso si fa riferimento ad un apprezzamento che prevede sensazioni corporee, preferenze, desideri e predisposizioni. Il compito del soggetto sperimentato è quindi quello di spostare il cursore verso destra o verso sinistra, scegliendo in sintesi se si trova più d'accordo a interpretare e descrivere l'azione del robot in termini mentalistici – adottando quindi l'atteggiamento intenzionale – o meccanici – propendendo per l'atteggiamento progettuale. Alla fine del test, per ognuno degli individui viene elaborato un *InStance Score* (ISS, Marchesi et al. 2019: 5), ovvero un indice numerico che rappresenta la tendenza del soggetto ad attribuire stati mentali o meno al robot, in una scala in cui 0 rappresenta la massima tendenza a scegliere descrizioni meccaniche e 100 la massima tendenza a scegliere descrizioni mentalistiche.

I risultati dello studio, in cui il test è stato sottoposto a 106 soggetti, mostrano una propensione generale a scegliere le descrizioni meccaniche a discapito di quelle mentalistiche, quindi non



adottando l'atteggiamento intenzionale nei confronti del robot. Infatti, la media degli ISS per la totalità degli ISQ somministrati è stata calcolata a 40,73, avvicinandosi quindi di più a una tendenza generale verso l'atteggiamento progettuale. Tuttavia, come sottolineato dagli autori, l'indice testimonia di una possibile, anche se meno adottata, tendenza dei soggetti a impiegare l'atteggiamento intenzionale, almeno in alcuni casi: «This shows that in principle it might be possible to induce adoption of the intentional stance toward artificial agents» (2019: 9). In sostanza, anche se l'esperimento fornisce evidenza che nella maggior parte dei casi i soggetti propendono per spiegazioni meccaniche del comportamento del robot, è però possibile adottare l'atteggiamento intenzionale di fronte ad agenti artificiali.

In generale, il test e l'esperimento che è servito a validarlo hanno il merito di puntare l'attenzione sui processi di attribuzione di intenzionalità ai robot sociali analizzando il rapporto che intercorre tra cognizione sociale e verbalizzazione dei processi di comprensione; inoltre, contribuiscono a spostare l'attenzione sui processi di attribuzione come correlati all'incontro con il robot, e non come esclusivamente dipendenti da meccanismi interni alla mente dell'essere umano. Il modo in cui questa analisi viene portata avanti, d'altra parte, presenta *in nuce* alcune questioni che potrebbero risultare problematiche, relative alle modalità di intendere l'intenzionalità e il processo che la vede impiegata nella cognizione sociale nell'interazione tra umani, che a sua volta si riverbera nella considerazione del robot come partner sociale e che abbiamo visto pervade tutto il campo di studi della HRI: il connubio con la ToM.

È interessante notare che, nell'articolo in oggetto, più volte gli autori ribadiscono una netta distinzione tra teoria dell'atteggiamento intenzionale e teoria della mente.

Although the two concepts are very tightly linked [...], they do differ with respect to the context in which they have been introduced in literature, and, what follows, in the empirical ways of addressing the concept. [...] if one has a ToM of another human's behavior, one has adopted the intentional stance, but not vice versa. [...] So I can fail the ToM test, but I can still adopt the intentional stance. In this way, with reference to human agents, the intentional stance is a necessary condition for ToM, but not a sufficient one. (Marchesi et al. 2019: 2).

In altre parole, in questo particolare studio, l'*intentional stance* e la ToM si riferiscono a due meccanismi complementari del processo di attribuzione di stati mentali: la prima, che indica la predisposizione dell'agente ad assumere una strategia di interpretazione basata sugli stati mentali del sistema osservato, consente alla seconda di realizzarsi, visto che la teoria della mente consiste nell'attribuzione di stati mentali, ma non ne garantisce la corretta ascrizione. Si potrebbe, quindi, sbagliare nell'attribuzione degli stati, ma la presenza della ToM implica che si debba adottare comunque l'atteggiamento intenzionale.

Come anche dimostrato nel caso della non necessità di un'intenzionalità genuina nel robot sottolineata in precedenza, gli autori non fondono completamente le due teorie, ma delineano un quadro in cui la teoria dell'atteggiamento intenzionale si coniuga attraverso gli strumenti concettuali della ToM senza necessariamente dividerne i presupposti. Tale posizione è rinforzata dal fatto che, anche qui a più riprese, gli autori dichiarano che lo scopo dell'articolo «is *not* to defend ToM accounts of social cognition» (Marchesi et al. 2019: 3, corsivi e grassetto degli autori). Tuttavia, gli stessi riconoscono che, sulla base di una letteratura specialistica sul tema (Baron-Cohen 1993; Leslie 1994), i due meccanismi possono essere tenuti insieme dal momento che entrambi fanno riferimento agli stati mentali per sviluppare il processo di comprensione dell'altro (Marchesi et al. 2019: 2), tanto da poter costituire un fronte comune contro le critiche sferrate da parte degli account interazionali della cognizione sociale e della *direct perception* come l'approccio enattivista. Ritorneremo sulla possibilità di evitare le critiche da parte di questi approcci nel paragrafo successivo. Per ora, dobbiamo registrare che in questo articolo – come anche in una serie di articoli precedenti e successivi scritti da componenti dello stesso gruppo di ricerca (Perez-Osorio & Wykowska 2019; Marchesi & Wykowska 2023) – la ToM e l'*intentional stance* sono considerate in rapporto strettissimo, soprattutto nel caso della cognizione umana negli adulti, ma sono ancora tenute separate, considerate come due strategie che si coniugano nel processo umano di comprensione intersoggettiva. Ulteriore prova del saldo legame tra le due è la somministrazione, insieme all'IST, di due test per la validazione della cognizione sociale e delle capacità di *mindreading*, utilizzati per sottoporre i soggetti sperimentati alla verifica delle capacità di lettura della mente e di attribuzione di stati mentali e per garantire che nessuno dei soggetti mostrasse difficoltà di comprensione intersoggettiva. In particolare, il *Reading the Mind in the Eyes test* (Baron-Cohen et al. 1999; 2001) consiste in una prova volta a valutare le abilità di inferire gli stati mentali dell'altro solamente guardando ai suoi occhi, mentre il test elaborato da Völlm et al. (2006) viene impiegato per valutare i correlati neurali della ToM e le capacità empatiche dei partecipanti all'esperimento, allo scopo di fornire una versione non verbale per la valutazione delle capacità di *mindreading* dell'individuo. Anche qui, le abilità di *mindreading* vengono considerate come un supporto alla attribuzione di intenzionalità, ma non perfettamente coincidenti: lo scopo dei ricercatori è quello di investigare forme generali di attribuzione di intenzionalità, non relative alla qualità degli stati mentali o alla loro correttezza<sup>137</sup>. Tuttavia, fin da qui è possibile notare una saldatura delle due visioni, in cui l'*intentional stance* è sì propedeutica al funzionamento della ToM, ma la ToM direziona l'attribuzione degli stati mentali e garantisce la

---

<sup>137</sup> «Please note that we focused on only one type of mental states (and, what follows, intentionality) attributed to artificial agents, namely the more explicit, propositional attitudes mental states carrying the inherent “aboutness” [...]. Our study did not address more implicit forms of adopting the intentional stance» (Marchesi et al. 2019: 3)

validazione della corretta attribuzione e la normatività circa l'adeguatezza dell'adozione della strategia.

Il rapporto tra *intentional stance* e ToM evolve in uno studio successivo, condotto qualche tempo dopo per validare i risultati di Marchesi et al. (2019) e introdurre misure implicite del fenomeno. In dettaglio, in Marchesi et al. (2021a), il gruppo di ricerca riconfigura l'impianto sperimentale allo scopo di testare se e in che misura ci sia una differenza nell'attribuzione di intenzionalità se il sistema osservato è un altro essere umano, Sara, o il robot iCub, mantenendo il IST come linea direttrice e dunque, almeno negli intenti, l'impianto teorico dell'*intentional stance* formulato da Dennett come punto di partenza.

Il nuovo scenario sperimentale prevede, in linea con l'IST, la scelta di una descrizione tra due, una che include attribuzione di stati mentali e l'altra con termini meccanici. A differenza dell'impianto classico dell'IST, d'altra parte, in questo esperimento i) al blocco di sequenze di azioni svolte dal robot viene aggiunto un secondo blocco in cui la stessa azione è svolta da un essere umano, ovvero Sara, nome che contiene lo stesso numero di lettere di iCub così da non condizionare l'individuo nel processo di scelta<sup>138</sup> e ii) la scelta non è basata sulle preferenze circa la descrizione mentalistica o meccanica, ma prevede che il soggetto dichiari il suo accordo o disaccordo con la descrizione proposta dallo studio, che può essere mentalistica o meccanica a prescindere dal personaggio che svolge la sequenza. Ogni sessione dell'esperimento, quindi, prevede uno scenario preso dall'IST originale, con iCub che svolge un'azione in una sequenza di tre fotografie, e una frase che descrive l'azione o in termini mentalistici (ad es. «iCub/Sara was trying to cheat by looking at opponent's cards») o meccanici (ad es. «iCub/Sara was unbalanced for a moment» (Marchesi et al. 2021a: 3)).

In questo task, oltre alla misura esplicita del questionario (*acceptance/rejection rate*), viene aggiunta una misurazione implicita relativa ai tempi di risposta (*time response*), allo scopo di includere una verifica su un livello "inconscio" della scelta operata.

Explicit attitudes operate on a conscious level and are generally measured through explicit self-reports (e.g. questionnaires), while implicit attitudes rely on unconscious and automatic processes, and are typically assessed via implicit measures (e.g. response time paradigms, implicit association test). Research suggests that implicit attitudes might constitute better predictors of future intentions and behaviors, and thus be more representative of *real attitudes* than explicit declarations. (Marchesi et al. 2021a: 334, corsivi nostri).

La misurazione della componente implicita del comportamento relativo all'attribuzione di intenzionalità, dunque, sembra una scelta dettata dalla necessità di rinforzare i risultati dello studio e

---

<sup>138</sup> Si ipotizza che tale scelta sia dovuta alla condizione in cui viene svolto lo studio, dato che, di norma, il compito prevede che il soggetto faccia la sua scelta nel modo più veloce possibile. Non si può tuttavia affermare con sicurezza, dato che lo studio riporta solamente la seguente specifica: «Note that the name of the character (iCub or Sara) contains an equal number of letters, in order to avoid biases during the reading process» (Marchesi et al. 2021a: 334-335).

fornire una base “naturale” che completi la rilevazione delle misure esplicite, affidata al questionario. In sostanza, lo scopo dello studio è quello di a) replicare e validare lo studio effettuato in Marchesi et al. (2019) e b) testare se i tempi di risposta dei partecipanti differiscono «while *processing* the two stances» (Marchesi et al. 2021a: 334, corsivi miei). Per quanto riguarda b), i ricercatori ipotizzano per i partecipanti allo studio avrebbero accettato più spesso (livello esplicito) e avrebbero avuto tempi di risposta più veloci (livello implicito) nei casi in cui la sequenza avesse rappresentato Sara compiere l’azione e la descrizione fosse stata in termini mentalistici, mentre i tempi di risposta più lenti (risposta implicita) e il maggior numero di disaccordi (risposta esplicita) si sarebbero riscontrati nel caso della descrizione mentalistica accoppiata con la sequenza in cui è rappresentato iCub.

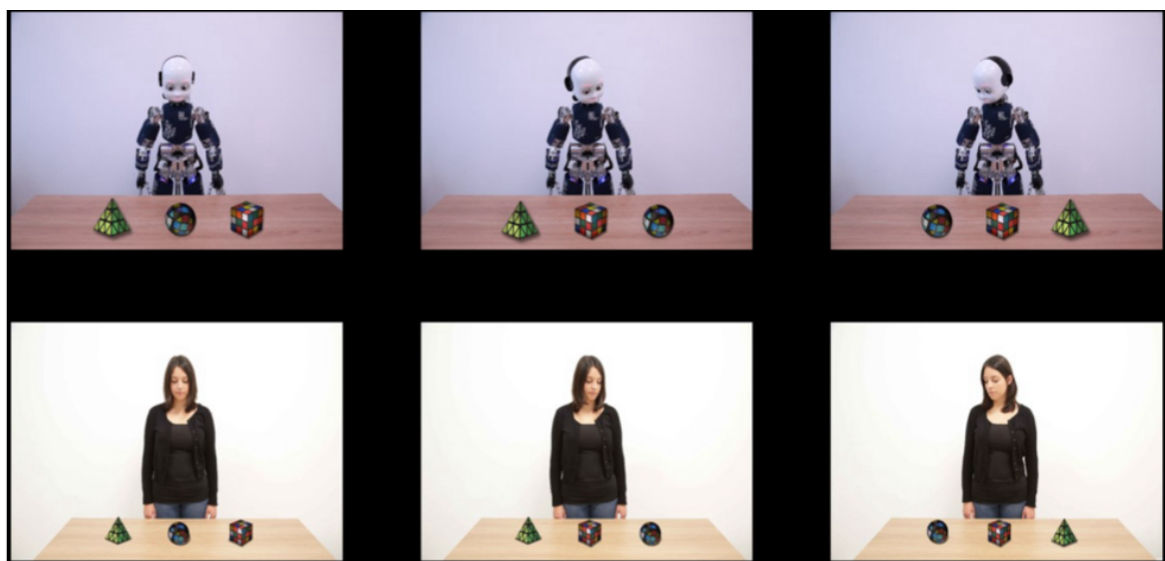


Fig.4.2 Schermate che rappresentano lo stesso scenario eseguito da iCub (sequenza superiore) e dall’essere umano Sara (sequenza inferiore). (Marchesi et al. 2021a)

Arrivando ai risultati, l’obiettivo a) risulta confermato in quanto a livello esplicito, attraverso il questionario, la maggior parte delle risposte per le sequenze con iCub ha registrato accordo con le descrizioni meccaniche e disaccordo con quelle mentalistiche, al contrario di quanto accaduto per lo scenario con Sara. Ancora più interessanti, però, sono i risultati relativi a b):

- i tempi di risposta più rapidi si sono registrati nella condizione “attore umano/descrizione mentalistica”, decisamente più veloci della condizione “attore umano/descrizione meccanica”, ma anche nel caso di “attore robotico/condizione mentalistica”. Allo stesso tempo, anche la condizione “attore robotico/descrizione meccanica” ha registrato tempi di risposta rapidi e più veloci rispetto alla condizione “attore umano/descrizione meccanica”. Il confronto tra i diversi tempi di risposta registrati nei diversi binomi fa propendere per un’interpretazione leggermente diversa rispetto a quella data in Marchesi et al. (2019): in questo caso, ciò che si registra è la *facilità* a adottare la strategia intenzionale di fronte ad esseri umani, seguita poi dalla

propensione a adottare l'atteggiamento progettuale di fronte al robot. In sostanza, se messi a confronto i due scenari con attori diversi, l'atteggiamento intenzionale di fronte all'essere umano è la scelta meno difficile da compiere.

- D'altra parte, prendendo in considerazione solamente lo scenario in cui è presente iCub e l'alternativa alle descrizioni tra mentalistico e meccanico, la situazione si fa molto più sfumata. Infatti, i ricercatori notano come i tempi di risposta, in questi casi, non siano così diversi.

Interestingly, when looking only at the robotic agent condition, no significant difference in the response times between mechanistic and mentalistic descriptions was observed. This is in contrast to the human condition, where the mentalistic descriptions were evaluated faster than mechanistic descriptions. (Marchesi et al. 2021a: 337).

In altre parole, la differenza che c'è tra i tempi di risposta nell'accettare o rifiutare la descrizione per lo scenario con il robot non sono paragonabili a quelli dimostrati nello scenario con l'essere umano. Secondo i ricercatori, ciò implica che, se per l'essere umano non ci sono dubbi su quale strategia adottare quando si cerca di descrivere l'azione osservata, per il robot potrebbe esserci qualche insicurezza, che impedisce al soggetto dell'esperimento di decidere facilmente e rapidamente il grado di accettazione nei confronti della descrizione proposta.

Ciò che risulta più interessante per noi è che, per fornire un'interpretazione circa le cause della differenza riscontrata, Marchesi e colleghi elaborano una lettura del meccanismo dell'*intentional stance* strettamente collegato alle modalità implicite di rilevazione e alle funzioni cerebrali credute il cuore del meccanismo della ToM:

This pattern suggests that the intentional stance might be the default strategy to explain other humans' behaviors both at the explicit and implicit level. When it comes to a robotic agent, participants might not find any of the descriptions as unlikely as the mechanistic descriptions for the human, meaning that both stances are compatible with participants' mindset, at least at the implicit level. However, at the explicit level, participants are more likely to accept mechanistic descriptions of the robot actions. This might reflect that both the intentional and design stance are strategies that could be adopted to explain an artificial agent's behavior. Some authors assume that the intentional stance would be the default stance in general (relying on the social cognition system) (Jack et al. 2013a; Schilbach et al. 2008). Indeed, neural areas (e.g., the superior temporal sulcus, lateral fusiform gyrus, medial prefrontal cortex, posterior cingulate, insula and amygdala) associated with mentalizing (*i.e.* the attribution of mental states to others), belong to the default mode network (Spreng & Andrews-Hanna 2015), suggesting that this key social cognition mechanism is the default mode of functioning of the brain (Spreng & Andrews-Hanna 2015; Jack et al. 2013b; Spreng et al. 2016). However, although a default system, it can be controllable (Urquiza-Hass & Kotrschal 2015), this would imply that at the earlier (implicit) stage of processing, individuals may adopt the intentional stance towards a robot but upon deliberate reflection, they would quickly identify the robot as an artifact, and they would reject the mentalistic descriptions. (Marchesi et al. 2021a: 337)

In altre parole, secondo l'interpretazione fornita dagli autori dello studio, l'*intentional stance* non sarebbe adottata a livello esplicito, momento "successivo" del processo inferenziale di comprensione

dell'altro misurato tramite l'IST, quanto invece a livello implicito, misurato tramite i tempi di risposta e localizzabile in una parte specifica del cervello, ovvero il *Default Mode Network* (DMN).

Nello specifico, gli studi di settore indentificano il DMN come «a neurocognitive brain network linked to various modes of self-generated thoughts» (Spreng & Andrews-Hanna 2015: 165) ed è strettamente legato negli studi sulla cognizione sociale al processo di *mentalizing*, ovvero «the cognitive process of inferring the thoughts, feelings, or beliefs of other people» (Ibid.). Il legame è reso necessario dal presupposto che «mentalizing [...] relies on self-generated cognition because humans do not have immediate perceptual access to other people's thoughts» (Ibid.), e quindi è necessario ipotizzare un meccanismo interno all'individuo che permetta di comprendere gli altri, invisibili nelle loro intenzioni, attraverso un ragionamento basato sui loro stati mentali. Il sodalizio tra *intentional stance* e ToM è dunque rinforzato non solo dagli studi sul funzionamento dell'area cerebrale che comprende i meccanismi di attribuzione di stati mentali agli altri, ma anche e soprattutto dal fatto che l'*intentional stance* sia associabile al DMN, come suggerito dagli studi citati nel passo riportato, e supporto neuronale della ToM: in questo modo, l'atteggiamento intenzionale diventa il correlato dell'attivazione dei meccanismi di attribuzione di stati mentali, riconosciuti come “*the default mode of functioning of the brain*”.

Sarà facile notare che, a supportare questa visione, interviene un'interpretazione in cui *intentional stance* e ToM non sono più solamente in rapporto complementare. Diversamente dallo studio precedente (Marchesi et al. 2019), in cui le due istanze per la comprensione dell'altro venivano tenute separate anche se strettamente legate, nel presente studio c'è una completa sovrapposizione dei due meccanismi: l'*intentional stance* – e la sua adozione – è considerata esattamente una manifestazione delle attività di attribuzione di stati mentali agli altri, il cui meccanismo interno è realizzato in specifiche parti del cervello in cui si produce l'inferenza adeguata a seconda della *stance* adottata. In questo modo, dunque, l'*intentional stance* trova le sue radici neurocognitive tramite il processo di *mentalizing*, e apre alla possibilità di interpretare i risultati dello studio come una dissonanza cognitiva: gli individui adottano l'*intentional stance* a livello implicito, ovvero a livello cerebrale, ma di fronte al robot, attraverso una deliberata riflessione, lo identificano come un artefatto, propendendo dunque per una descrizione secondo la *design stance* e respingendo le interpretazioni mentalistiche.

Ora, la visione proposta in questo studio presenta secondo noi diverse problematiche, che sono indicative del modo in cui l'attribuzione di intenzionalità è considerata negli studi attuali in HRI e che include una interpretazione particolare dell'*intentional stance* legata alla ToM e al processo di mentalizzazione. In primo luogo, il modo in cui il DMN viene richiamato a supporto della localizzazione dell'atteggiamento intenzionale in pattern neuronali si basa sul presupposto che questo sia “*the default mode of functioning of the brain*”. In realtà, il termine *default* nella sigla DMN non

sta ad indicare una modalità di funzionamento del cervello standard, come è invece assunto in questo studio, ma al fatto che il DMN è stato scoperto essere attivo quando il cervello non è impegnato in compiti cognitivi esterni specifici, ovvero quando è uno stato di “riposo”: uno stato di *default* (Andrews-Hanna 2012). In altre parole, il DMN è attivo quando una persona non sta svolgendo attività focalizzate come risolvere problemi o interagire con l’ambiente, ma piuttosto quando è assorta nei suoi pensieri, in attività spesso definite di *mindwandering* (Corballis 2015) come la riflessione su sé stessi, la memoria episodica o l’immaginazione del futuro. Anche se le correlazioni tra DMN e cognizione sociale sono inizialmente state istituite sotto il concetto di mentalizzazione (Spreng & Andrews-Hanna 2015), il legame con la teoria della mente come processo di attribuzione di stati mentali agli altri che segue logiche strettamente razionali è stato messo in dubbio in favore di una visione più ecologica e situata dell’individuo e della sua cognizione in contesti co-partecipati (Yeshurun et al. 2021; Lobaccaro 2022). Ciò che vogliamo sottolineare, in questo caso, è che la prospettiva fondazionale dalla ToM, in cui l’individuo è un elemento solitario che elucubra pensieri interni al fine di decifrare gli stati mentali dell’altro, ha generato alcune conseguenze a cascata: i) ha condotto a poter allacciare il meccanismo di attribuzione di stati mentali al funzionamento del cervello quando è impegnato in pensieri ego-riferiti e che non prevedono il coinvolgimento di molti stimoli esterni; ii) la sovrapposizione della mentalizzazione con l’*intentional stance* ha portato i ricercatori a ipotizzare che una misurazione implicita dei tempi di risposta potesse rivelare un’adozione della strategia a livelli più “profondi”, e quindi a rintracciare una base neurocognitiva da testare per validare l’adesione a una o l’altra delle prospettive interpretative fornite dall’approccio dennettiano all’intenzionalità, considerate anti-correlate nell’attivazione di differenti pattern neurocognitivi (Jack et al. 2013b).

Inoltre, i ricercatori suffragano l’ipotesi che l’*intentional stance* sia individuabile in alcuni pattern neuronali, in particolare afferenti al DMN, richiamando alcuni studi di settore che già hanno provato una correlazione di questo tipo. Nello specifico, lo studio di Jack et al. (2013a) citato in Marchesi et al. (2021a) fa riferimento a una serie di studi sulla deumanizzazione<sup>139</sup> in cui si registra un’anti-correlazione nell’attivazione di meccanismi neurocognitivi quando l’individuo ragiona circa stati mentali interni (DMN attivato/TPN disattivato) e quando ragiona riguardo processi di natura meccanica (DMN disattivato/TPN attivato), fornendo una base neurobiologica per la teoria dei due tipi di deumanizzazione proposta da Haslam (2006). Il ruolo dell’*intentional stance* è qui particolare, dal momento che «enables us to make predictions about others’ actions and causes us to adopt a more emotionally detached view of them» (Jack et al. 2013a: 314). Il fatto è che, però, l’*intentional stance*

---

<sup>139</sup> Il richiamo a questo tipo di studi corrobora ulteriormente la tesi espressa in §3.4.1 riguardo il legame istituito tra ToM e studi sulla deumanizzazione, in cui quest’ultima è messa a confronto con i meccanismi creduti responsabili dell’antropomorfismo.

chiamata in causa non è quella teorizzata da Dennett (1971; 1986), ovvero una strategia che si avvale del linguaggio mentalistico per interpretare e prevedere il comportamento del sistema osservato, ma una modalità cognitiva che si origina dalla fusione di altre due modalità, ovvero la *physical stance* – una modalità cognitiva responsabile del ragionamento analitico come quello scientifico e logico, attività registrata all’attivazione del Task Positive Network (TPN) – e la *phenomenal stance* – ovvero la modalità cognitiva attraverso cui si ragiona sulle esperienze e sulle preoccupazioni di ordine morale, afferenti al DMN (Jack et al. 2013a: 314). In questo senso, l’*intentional stance* è considerata la modalità cognitiva che permette il ragionamento sociale affettivamente distaccato, in cui aspetti di entrambi i network si attivano. Si può notare, dunque, che, nonostante la chiara omonimia, i concetti implicati sono alquanto diversi, tanto che nello studio non c’è alcuna traccia di testi di Daniel Dennett, né riferimenti alla sua teoria degli atteggiamenti. Lo stesso vale per un altro studio degli stessi autori, citato in Marchesi et al. (2021a), in cui la *phenomenal stance* è inoltre identificata come «the ‘default mode’ of social cognition» (Jack et al. 2013b: 396). Anche in Spreng et al. (2016) e in Schilbach et al. (2008), l’*intentional stance* non è menzionata, ma ci sono alcune evidenze che è interessante sottolineare, che non sono le stesse illuminate dallo studio di Marchesi et al. (2021a). In particolare, Schilbach et al. (2008) riportano che, date numerose evidenze circa la sovrapposizione di “resting state” a livello di pattern neuronali e compiti di cognizione sociale, una delle attività di *default* dell’essere umano sia rintracciabile nella propensione a entrare in dinamiche sociali. Non è, dunque, l’*intentional stance* di per sé a costituire la modalità standard di funzionamento del cervello; è la cognizione sociale, però, a definirsi come modalità pervasiva e di *default* di comprendere, «to which [human beings] return when not faced with a given task» (Schilbach et al. 2008: 464).

In ultima analisi, il quadro teorico su cui si basa l’interpretazione delle misure esplicite ed implicite in Marchesi et al. (2021a) è supportato da una visione in cui all’esposizione allo stimolo visivo che rappresenta il robot svolgere l’azione corrisponde dapprima una reazione neurocognitiva, in cui l’*intentional stance* è adottata e riscontrabile nei tempi di risposta rapidi per il binomio “individuo/descrizione mentalistica”, ma anche nei livelli comparabili tra i tempi di risposta nei binomi “robot/descrizione mentalistica” e “robot/descrizione meccanica”. Allo stesso tempo, la propensione ad accettare descrizioni meccaniche per i robot sarebbe maggiore perché, successivamente, riflettendo su cosa scegliere, l’individuo non troverebbe conferma nelle proprie credenze circa la plausibilità di una descrizione mentalistica per le azioni del robot, riconoscendo invece lo stesso come un artefatto e optando quindi più facilmente per l’accordo con le proposizioni in termini meccanici. In ogni caso, quello descritto è un processo in cui due diversi meccanismi sono in gioco, ma non allo stesso tempo: la risposta implicita, precedente, pone le basi per un certo tipo di riconoscimento dello stimolo in entrata, che però viene in questo caso contrastata da una riflessione,



successiva, circa lo statuto dell'oggetto riconosciuto. Come la tradizione neo-cartesiana suggerisce, dunque, il processo di attribuzione di intenzionalità passa per un'inferenza alla miglior spiegazione sugli stati mentali che causano il comportamento osservato, in cui gli input percettivi passano al vaglio del ragionamento deliberato prima di produrre un output interpretativo. In questo aspetto si intravede come la posizione cognitivista si staglia sull'impostazione neo-comportamentista rispetto all'intenzionalità, facendo sbilanciare l'approccio verso una visione internalista e computazionale della mente: termini come «*processing the two stances*» (Marchesi et al. 2021a: 334, corsivi nostri) o «*real attitudes*» (ibidem, corsivi nostri) riferito alle evidenze implicite – considerate la vera essenza del comportamento, interne alla mente dell'individuo e rilevabili solamente tramite processi inferenziali – rivelano una propensione all'interpretazione dell'esperimento utilizzando le terminologie e le posizioni teoriche proprie non dell'approccio neo-comportamentista, quanto invece a quello internalista del neo-cartesianesimo.

Tuttavia, come già evidenziato in precedenza, il ragionamento deduttivo tipico dell'approccio neo-cartesiano all'intenzionalità mal si accoppia con il procedimento induttivo alla base dell'account neo-comportamentista, di cui Dennett è il primo sostenitore. Vedremo a breve e in dettaglio le inconsistenze che rendono molto difficile la possibile coesistenza di questi due approcci in una visione unitaria, soprattutto se il primo è concepito nella sua variante neurocognitivamente situata; per ora, è sufficiente ricordare che esiste un'incompatibilità di fondo tra i due, che si può superare solamente a scapito della solidità degli approcci chiamati in causa.

Un'ultima problematica che vogliamo sottolineare rispetto all'impianto sperimentale dell'IST, sia nella modalità originale (Marchesi et al. 2019), sia nella sua forma complessa di integrazione tra misurazione esplicita ed implicita (Marchesi et al. 2021a) è relativa alla categorizzazione dei termini che formano le descrizioni proposte ai soggetti dell'esperimento. Per ovviare alle critiche che avrebbero potuto generarsi se la scelta dei termini fosse stata arbitraria, nel primo studio i ricercatori hanno sottoposto la lista delle descrizioni a una validazione esterna: un gruppo di controllo aveva il compito di valutare solamente le frasi (senza accoppiamento con le sequenze di immagini) per come venivano comprese, seguendo una *Likert scale* di 10 punti in cui a 0 corrispondeva il “completamente meccanico” e al 10 il “completamente mentalistico” (ivi: 4). Anche se i risultati del test hanno riportato soglie adeguate per l'impiego nell'IST, quello che vogliamo sottolineare è che il compito di giudicare l'indice meccanico o mentalistico di un termini non sembra affatto scontato: ad esempio, nel materiale supplementare allo studio<sup>140</sup>, coppie di descrizioni come “iCub aligns the head with the webcam/iCub is surprised by the webcam moving”, “iCub orders Italian cities alphabetically/iCub prefers Italian cities to other cities”, “iCub easily detects objects with screens/iCub finds digital

---

<sup>140</sup> <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00450/full#supplementary-material>

technology intriguing” non sembrano essere mutualmente esclusive, né denotare massimamente l’uno o l’altro atteggiamento. Questo fattore è testimoniato anche dalla valutazione finale delle descrizioni, che si attesta su 8.2 per l’atteggiamento intenzionale e su 4.3 per quello meccanico: il fatto è che, anche se l’individuo compisse una scelta più incline alla descrizione meccanica, come per “iCub orders Italian cities alphabetically”, l’atteggiamento intenzionale non sembra essere completamente escluso, poiché si potrebbe interpretare la frase attribuendola a un essere umano senza per forza implicare che sia l’atteggiamento progettuale ad essere adottato. Questo elemento si riverbera, di conseguenza, anche nello studio di Marchesi et al. (2021a), anche se nelle differenti modalità di somministrazione che abbiamo visto in precedenza.

C’è da illuminare, però, un aspetto positivo messo in risalto dagli studi che abbiamo analizzato fin qui. È indubbio, infatti, che sia possibile registrare risultati discordanti sull’adozione dell’*intentional stance* se misurata attraverso modalità esplicite e implicite. Questo ha portato gli studi a ipotizzare che i processi cognitivi responsabili dell’attribuzione di intenzionalità ai robot sociali avvengano a livelli distinti e, a volte, completamente autonomi l’uno dall’altro, come sottolineato da denotazioni quali *low-road* vs. *high-road processing* (Fiala et al. 2014), *implicit* vs. *explicit processing* (Banks 2020; Takahashi et al. 2013), *first-line* vs. *second-line reasoning* (Levin et al. 2012), o *type1* vs. *type2 processing* (Zlotowski et al. 2018), raccolti sotto l’etichetta di *dual processes theories* (Evans 2003). Come notato anche da Thellman et al. (2022), la maggior parte degli studi che si occupa di indagare il fenomeno ha ipotizzato che l’attribuzione di stati mentali ai robot sia un processo in due fasi, una basata su processi di pensiero che avvengono ad un alto livello cognitivo (di tipo proposizionale, logico e inferenziale) e l’altra afferente a processi cognitivi di basso livello (ovvero corporei e impliciti, compresi quelli neurocognitivi). Il disaccordo, però, sta su quale venga prima: negli studi che abbiamo analizzato, il processo di deliberata decisione misurato tramite questionario è considerato essere successivo all’adozione dell’*intentional stance* in modalità implicita, mentre in altri studi (Martini et al. 2016; Wiese et al. 2017; Wykowska et al. 2014) l’ordine è capovolto, descrivendo un processo in cui «the individual must engage in a higher-level thought process in which he or she decides to adopt an interpretative mentalistic or intentional stance toward the robot *before* lower-level attribution processes are activated» (Thellman et al. 2022: 12).

Il quadro che si delinea, allora, ci conduce a dover approfondire il modo in cui l’adozione di un certo tipo di teoria riguardo all’intenzionalità e alla sua attribuzione nei confronti degli altri condizioni alcune scelte interpretative e ne precluda delle altre, arrivando a comporre un insieme di problematiche che sembrano dover essere affrontate sotto l’egida della ToM. Cercheremo di mostrare, tuttavia, che le domande poste sono connotate da una obbligatorietà di soluzione solamente apparente,

che può essere svelata se si adotta una diversa concezione di intenzionalità e del suo ruolo nella cognizione sociale.

#### 4.2.3 *Vecchie e nuove ibridazioni*

Arrivati qui, appare chiaro che l'ibridazione concettuale tra i due approcci sviluppati in merito all'intenzionalità nelle scienze cognitive mostra i primi segni di una frattura dovuta ai disequilibri innescati dallo sbilanciamento verso la visione internalista e modulare della mente e della cognizione. Infatti, abbiamo visto che il binomio tra approccio neo-cartesiano e neo-comportamentista all'intenzionalità, nonostante i diffusi esempi di integrazione tentati nelle scienze cognitive (Leslie 1994; Baron-Cohen 1991, 1993), comporta diverse rinunce da parte dell'una o dell'altra concezione. In questo specifico caso, il problema si pone nel momento in cui gli autori dello studio (Marchesi et al. 2021a) chiamano in causa esattamente le posizioni di Dennett sull'atteggiamento intenzionale per impostare il discorso sull'attribuzione di intenzionalità ai robot sociali e per costruire il test che prevede misure implicite ed esplicite, appoggiandosi alle riflessioni sviluppate nello studio precedente, ma interpretano i risultati impliciti attraverso uno sguardo riduzionista che si affida completamente ai dettami della teoria della mente. Sicuramente, l'adesione preliminare all'approccio dell'atteggiamento intenzionale è dovuta – almeno in parte – alla volontà di replicare lo studio di Marchesi et al. (2019), il quale però non comportava problemi relativi all'integrazione di misure implicite e anzi si sviluppava tutto su una dimensione linguistica di validazione. I problemi aperti in Marchesi et al. (2021a) sono relativi alla volontà di localizzare l'attribuzione di intenzionalità in meccanismi cerebrali e ciò conduce a stringere il nodo tra *intentional stance* e ToM fino a far collassare la prima sulla seconda.

Il motivo per cui si creano le difficoltà che abbiamo segnalato sul versante teorico relativo alle nozioni di intenzionalità implicate nello sguardo adottato in questo approccio è relativo alla difficoltà di naturalizzare la nozione di atteggiamento intenzionale per come proposta originariamente da Dennett (1971, 1987). Infatti, il filosofo sviluppa la sua teoria degli atteggiamenti proprio in aperto contrasto con le teorie computazionaliste e funzionaliste più in voga in quegli anni, rappresentate dalle posizioni di Jerry Fodor (1975, 1983), dando via a una corrente alternativa di stampo finzionalista. In particolare, le differenze che portano a due visioni incompatibili dell'intenzionalità riguardano almeno due questioni. La prima è relativa alla considerazione della struttura della mente: se per la visione fodoriana la mente è modulare, ovvero composta da diversi sistemi cognitivi indipendenti l'uno dall'altro (*incapsulati*) e specializzati in compiti precisi (Fodor 1983), per Dennett la mente si configura come emergente da dinamiche complesse di elaborazione distribuita, non necessariamente coincidenti con porzioni del cervello che dovrebbero essere in rapporto diretto a stati

mentali definiti (Dennett 1991). In secondo luogo, la differenza tra i due è rintracciabile nel disaccordo circa il funzionamento della mente: la visione radicalmente computazionalista di Fodor (1975), secondo cui la cognizione è essenzialmente una forma di computazione simbolica, viene contrastata da Dennett con un approccio che rifiuta la riduzione a macchina simbolica e delineando una concezione di mente come prodotto dei processi evolutivi e di adattamento (Dennett 1987, 1991), seguendo quindi una concezione darwinista e aprendo la strada alla considerazione del ruolo della biologia nei processi cognitivi anche nella corrente *mainstream* delle scienze della mente (Dennett 1995, 2017).

La sostanziale differenza di visioni riguardo la struttura e il funzionamento della mente non poteva che scaturire nella concezione antitetica nei confronti dell'intenzionalità: mentre per Fodor (1975) il significato è determinato dal contenuto delle rappresentazioni mentali, essendo la mente una macchina che manipola rappresentazioni simboliche della realtà attraverso un linguaggio del pensiero, e l'intenzionalità è intrinseca alle rappresentazioni, per Dennett (1987) quest'ultima si realizza nei modi in cui ci relazioniamo alle rappresentazioni, attribuendo quindi intenzionalità ai sistemi che osserviamo. Il concetto di *intentional stance*, insieme agli altri atteggiamenti adottati per predire e interpretare il comportamento attenzionato, è uno dei contributi chiave derivati da questo approccio, in cui attribuiamo intenzionalità a un agente per comprendere e predire il suo comportamento senza assumere che l'agente possieda davvero stati mentali interni di tipo simbolico-rappresentazionale.

Diventa molto complicato, dunque, se non impossibile, riuscire a naturalizzare l'adozione dell'*intentional stance*, o di qualsiasi altra *stance*, rendendola il correlato linguistico di stati mentali interni, e ancor più complicato se tali stati mentali vengono pensati risiedere in – o essere riducibili a – configurazioni neurali deputate ad un unico scopo. Lo stesso Dennett aveva intravisto il rischio che il proprio impianto teorico fosse passibile di essere interpretato come aderente alla concezione computazionalista delineata da Fodor e ha ben spiegato l'incompatibilità con quelle posizioni in un celebre passo che chiarisce definitivamente la questione:

Jacques colpisce a morte suo zio in Trafalgar Square e viene arrestato sul posto da Sherlock; Tom legge questa notizia sul «Guardian» e Boris ne viene a conoscenza tramite «Pravda». Ora, Jacques, Sherlock, Tom e Boris hanno avuto esperienze considerevolmente differenti – per non parlare delle loro precedenti biografie e delle loro prospettive future – , ma c'è una cosa che hanno in comune: credono tutti che un francese abbia commesso un omicidio a Trafalgar Square. Nessuno di loro *dice* questo, neppure «a sé stesso»; *quella proposizione*, possiamo ipotizzare, non «occorre» in nessuno di loro, e anche se lo facesse, avrebbe un importo completamente diverso per Jacques, Sherlock, Tom e Boris. E tuttavia tutti credono che un francese abbia commesso un omicidio in Trafalgar Square. Questa è una proprietà comune che è visibile, per così dire, solo da un punto di vista molto limitato – il punto di vista della psicologia del senso comune. Gli psicologi del senso comune ordinari non hanno alcuna difficoltà ad attribuire queste utili, ma sfuggenti comunanze alle persone. *Se dunque sostengono che nel fare così devono postulare un oggetto strutturato*

*in modo analogo nella testa di ciascuno, ciò rappresenta un bel saggio gratuito di concretezza fuori luogo, una deplorabile caduta nell'ideologia.* (Dennett 1987: 83, corsivi miei).

Quindi, l'atteggiamento intenzionale non si può naturalizzare né in senso riduzionistico, né nel senso di correlati computazionali all'interno della mente, a meno di snaturare allo stesso tempo il concetto per come elaborato da Dennett. Questo perché, a differenza della reinterpretazione fornita negli studi successivi sia nelle scienze cognitive (Frith & Frith 2005; Gopnick & Wellman 1994), sia negli studi di HRI (Breazeal 2003a; Fong et al. 2003; Dautenhahn 2007a), l'atteggiamento intenzionale non è causato dall'individuazione di stati mentali che inducono l'adozione di una *stance* rispetto alle altre, ma è una *strategia* attraverso cui l'individuo cerca di capire, interpretare, spiegare e predire il comportamento del sistema che osserva producendo descrizioni che scommette riescano nel compito previsto. Non ci sono motori causali di generazione dell'atteggiamento, gli stati mentali non vengono prima percepiti dall'individuo e poi utilizzati per elaborare la strategia: esattamente al contrario, invece, gli stati mentali vengono attribuiti al sistema a partire dall'osservazione del suo comportamento e supportati nella validazione dall'efficacia della strategia nel predire le azioni future. Nel senso proposto da Dennett, tutte le *stance* sono degli strumenti realizzati dal linguaggio che sono utili per muoversi nell'ambiente di riferimento, ma non hanno niente a che fare con il *mentalese*. Ancora più chiaramente, l'*intentional stance* è composta di *abstracta*, ovvero costrutti logici che fanno riferimento alla psicologia del senso comune<sup>141</sup>:

However rational we are, it is the myth of our rational agenthood that structures and organizes our attributions of belief and desire to others and that regulates our own deliberations and investigations. We aspire to rationality, and without the myth of our rationality the concepts of belief and desire would be uprooted. Folk psychology, then, is *idealized* in that it produces its predictions and explanations by calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do. Folk psychology is *abstract* in that the beliefs and desires it attributes are not – or need to be – presumed to be intervening distinguishable states in an internal behavior-causing system. (Dennett 1987: 52, versione inglese)<sup>142</sup>

È questo il presupposto su cui si basa il principio del *true believer*, ovvero il fatto che l'atteggiamento intenzionale possa essere adottato anche per predire il comportamento di un sistema

---

<sup>141</sup> La composizione degli *abstracta* per le altre *stance*, invece, fa riferimento a entità legate alle leggi fisiche e matematiche (cfr. Dennett 1987: 51).

<sup>142</sup> Si è preferito riportare il testo nell'originale inglese, dal momento che la traduzione italiana include termini che, a opinione di chi scrive, connotano il passaggio di un'aura troppo idealista, slegata dalle dinamiche pratiche dell'interpretazione, che dipinge l'impostazione filosofica dennettiana sotto una luce internalista che mal gli si addice. Riportiamo comunque qui la traduzione in italiano, per rendere chiara la scelta operata: «Per quanto razionali siamo, è il mito del nostro essere agenti razionali che struttura e organizza le nostre attribuzioni agli altri di credenza e desiderio e che governa le nostre stesse deliberazioni e indagini. Noi aspiriamo alla razionalità, e senza il mito della nostra razionalità i concetti di credenza e desiderio sarebbero privi di radici. La psicologia del senso comune, per tanto, è *idealizzata* in quanto formula le sue previsioni e spiegazioni calcolandole entro un sistema normativo; essa prevede che cosa crederemo, desidereremo, faremo, mediante la determinazione di che cosa dovremmo credere, desiderare e fare. La psicologia del senso comune è *astratta* perché non si presume che le credenze e i desideri da essa attribuiti siano – o debbano essere – stati distinguibili che occorrono in un sistema interno di causazione comportamentale» (Dennett 1987: 78).

che genuinamente intenzionale non è. Nel pensiero di Dennett, in cui l'intenzionalità è quel legame che scaturisce in relazione alle modalità di rappresentazione del mondo – ma non è a loro intrinseca –, la normatività dell'atteggiamento è dettata dalle contingenze relazionali in cui i sistemi si incontrano, e non in relazione a una composizione interna che da sola dovrebbe rendere possibile l'emersione del significato.

Non è che noi attribuiamo (o dovremmo attribuire) credenze e desideri soltanto a cose in cui troviamo rappresentazioni interne, ma piuttosto che quando scopriamo un oggetto per il quale funziona la strategia intenzionale, cerchiamo di interpretare alcuni dei suoi stati o processi interni come rappresentazioni interne. Ciò che fa di una caratteristica interna di una cosa una rappresentazione potrebbe essere solamente il suo ruolo nel regolare il comportamento di un sistema intenzionale. (Dennett 1987: 53).

In sostanza, il punto su cui Dennett insiste è esattamente relativo alla questione della corrispondenza tra adozione dell'*intentional stance* tramite l'uso di termini mentalistici nella descrizione del comportamento del sistema e stati mentali che dovrebbero mostrare una precisa e distinta configurazione sul piano fisico. Facendo ancora appello alle parole di Dennett, la sua posizione di fa cristallina:

Aspettarsi un distinto meccanismo fisico dietro ogni genuinamente distinto stato mentale è una cosa; aspettarsi un meccanismo distinto per ogni pretesa distinzione che può essere espressa nel tradizionale linguaggio mentalistico è un'altra. (Dennett 1987: 62)

Ora, nella diffusamente adottata ibridazione con l'approccio neo-cartesiano all'intenzionalità che dà vita alla teoria dell'attribuzione degli stati mentali, è proprio questa la caratteristica a cui la posizione neo-comportamentista deve rinunciare, abdicando in favore di assunti ontologici internalisti che vedono gli stati mentali come ciò che causa effettivamente il comportamento e il processo di comprensione dell'altro come il ragionamento che si origina a partire dal legame causale con il comportamento che producono. Come più volte abbiamo sottolineato, la teoria delle *stance* di Dennett, invece, è portatrice di un approccio molto più externalista all'intenzionalità e non prevede alcuno specifico impegno nei confronti della conformazione interna della mente o degli stati mentali. Tutto ciò che conta è che il sistema agisca in accordo alla strategia intenzionale, il comportamento del quale non è necessariamente realizzato con il supporto di una intenzionalità originale. Il comportamento del computer, dell'animale, del cacciatore possono essere tutti descritti tramite l'atteggiamento intenzionale, non perché possono in qualche modo averla o simularla internamente, ma perché è il loro atteggiamento che si dispiega seguendo una direzionalità e suggerisce all'individuo di comprenderlo in termini mentalistici. È in questo legame istituito tra occhio osservatore e sistema osservato che si manifesta l'intenzionalità e l'atteggiamento intenzionale viene assunto perché, semplicemente, *funziona*: il vaglio dell'evoluzione ha selezionato questa come strategia efficace di predizione degli esseri viventi, un meccanismo inferenziale di tipo induttivo che

ci garantisce di prevedere gli scenari futuri e che si basa sul principio della migliore efficienza del sistema. In sostanza, è meglio attribuire troppa intenzionalità, che poca: dal punto di vista evolutivo, meglio sopravvalutare i sistemi agentivi, piuttosto che esporsi al rischio di essere presi di sorpresa.

La mia tesi è che l'atteggiamento intenzionale offra un punto di vista che consente di discernere modelli [utili]. Questi modelli sono oggettivi – sono *lì* per essere scoperti – ma, secondo il nostro punto di vista, non esistono *lì fuori* indipendenti da noi, poiché sono modelli composti in parte delle nostre stesse reazioni “soggettive” a ciò che è *lì fuori*; sono i modelli creati per dare un ordine ai nostri affari narcisistici. È facile per noi, fatti come siamo, percepire i modelli che sono visibili assumendo l'atteggiamento intenzionale – e solo questo atteggiamento. (Dennett 1987: 61)

Resta da chiarire un punto: se l'intenzionalità e la capacità di attribuirle agli altri non è nel cervello, se non è identificabile con una serie di stati fisici della materia che produce la mente, dunque, come è possibile afferrare l'atteggiamento intenzionale per testare se e in che modo è effettivamente presente nell'interpretazione di un sistema? Anche in questo caso, la risposta ci viene direttamente da Dennett. In una nota a margine, Dennett commenta un passo di Elizabeth Anscombe, in cui la filosofa descrive la capacità di comprendere gli altri come «un ordine che si dà ogni qualvolta le azioni sono compiute con le intenzioni» (Anscombe 1957: 80), dicendo:

«[Anscombe] non disse *in quale parte* del mondo questo ordine doveva essere percepito. Nel cervello? Nel comportamento? Per anni non sono riuscito a capire il senso di quest'affermazione, ma ora mi è chiaro che cosa la Anscombe potesse aver inteso e perché fosse così evasiva nella descrizione (e localizzazione) dell'ordine. Dire dove sia l'ordine intenzionale è tanto difficile quanto dire dove siano i modelli intenzionali nel mondo della Vita. Se “guardiamo” il mondo nel modo giusto, i modelli sono ovvi. Se guardiamo (o descriviamo) il mondo in un qualsiasi altro modo, essi sono, in generale, invisibili». (Dennett 1987: 61, nota 7)

È difficile, quindi, riuscire a capire in che luogo possa essere rintracciato il *dove* dell'intenzionalità, tanto più quando si tratta di comprendere altri sistemi, che in potenza potrebbero essere solamente prevedibili con l'atteggiamento intenzionale ma non possedere alcun tipo di intenzionalità – una sorta di *zombi* robotico, per come lo definirebbero sia Dennett (1991) che Chalmers (1996). Per ovviare al problema, Dennett elabora un approccio *eterofenomenologico* ai contesti sperimentali i cui i soggetti umani che devono esprimere giudizi linguistici in merito a delle proprie esperienze. L'approccio consiste nell'interpretare «le [loro] affermazioni come espressioni delle [loro] credenze» (1991: 110), ma senza la possibilità di sapere se il nesso sia garantito, né se quanto espresso a livello linguistico corrisponda effettivamente a una descrizione dell'esperienza provata<sup>143</sup>. In sostanza,

---

<sup>143</sup> «Come abbiamo già visto, le persone in alcune circostanze si sbagliano su ciò che stanno facendo e su perché lo stanno facendo. Non è esatto dire che stanno mentendo nella situazione sperimentale, è vero piuttosto che si lasciano andare alla confabulazione; riempiono i vuoti, immaginano, ipotizzano, scambiano il teorizzare con l'osservare. La relazione tra ciò che dicono e ciò che li spinge a dire ciò che dicono non potrebbe essere più oscura, sia a noi eterofenomenologi all'esterno sia ai soggetti stessi. Essi non hanno nessuno strumento per «vedere» (presumibilmente, con un occhio interiore) il processo che governa le loro asserzioni, ma ciò non impedisce loro di avere delle opinioni granitiche da esprimere» (Dennett 1991: 111).

l'eterofenomenologia viene pensata da Dennett come un metodo per interpretare un comportamento sulla base dell'atteggiamento intenzionale, che somiglia molto al modo in cui interpretiamo le opere di narrativa (1991: 89-96) e che consente di accedere attraverso un 'ponte' linguistico all'esperienza vissuta del sistema, costruendo un legame tra l'esperienza del soggetto sperimentato e il suo modo di esprimerla nell'esperimento che vuole testarla.

Vedremo più avanti (§4.3.3) come il legame con le opere di narrativa possa aiutarci a capire meglio come validare la presenza dell'intenzionalità nelle interazioni tra umani e robot. Ora, desideriamo esplorare alcune conseguenze che scaturiscono dalla rilettura dell'ibridazione degli approcci neo-cartesiano e neo-comportamentista all'intenzionalità, al fine di rilanciare il discorso e fornire una nuova possibile interpretazione del ruolo che l'intenzionalità ha nella HRI e del modo in cui sia possibile concepirla.

- Da un lato, attraverso l'analisi dell'impianto sperimentale di Marchesi e colleghi (2019, 2021a) abbiamo compreso che concepire l'attribuzione di intenzionalità come un meccanismo inferenziale di tipo deduttivo porta a un'errata applicazione e interpretazione del concetto di *intentional stance* così come sviluppato originariamente, e il tentativo di trovare modi impliciti per validarlo, sbilanciandosi verso un approccio internalista neo-cartesiano all'intenzionalità, conduce a generalizzazioni in termini di ToM e alle sue varianti che situano il meccanismo di comprensione dell'altro in aree specifiche del cervello deputate a questo esclusivo compito. I due approcci, quindi, anche se opportunamente ibridati, continuano a presentare delle difficoltà a integrarsi e comportano problemi sostanziali quando si arriva al momento di interpretare i risultati degli esperimenti svolti. L'incompatibilità delle idee che sta alla base di questa ibridazione – oramai di vecchia data – si è risolta con una svalutazione dell'approccio neo-comportamentista dennettiano in favore di uno sguardo internalista e modularista dei processi cognitivi. L'approccio dennettiano, però, mostra dei vantaggi rispetto a quello neo-cartesiano: nel campo della teoria generale, consente di mantenere una visione meno radicale sul solipsismo della mente e non riduzionista dei processi cognitivi, evitando di ricorrere a modelli modulari rigidi e internalisti che localizzano l'intenzionalità in specifiche aree cerebrali la cui localizzazione non ammette sfumature; per gli HRI, offre il vantaggio di non richiedere necessariamente che i robot siano dotati di stati mentali genuini o che sorpassino una soglia di capacità computazionale, ma che siano progettati in modo da esibire comportamenti interpretabili come intenzionali dall'osservatore umano. L'operazione che proponiamo, allora, è quella di restaurare l'integrità dell'approccio di Dennett all'intenzionalità, cercando di mantenere intatti i suoi principi epistemologici e, in linea con le posizioni espresse in §3, liberare l'approccio dalle contaminazioni con la teoria della mente e rendere possibile la



costruzione di un approccio all'intenzionalità basato sull'interazione. Ciò ci consentirà anche di sondare diversi metodi per testare empiricamente la presenza dell'intenzionalità nelle interazioni con i robot, prendendo in considerazione forme di validazione non solamente neurocognitiva, ma anche comportamentale.

- Dall'altro lato, dalla nostra analisi risulta che, se accettiamo la concezione dennettiana dell'atteggiamento intenzionale come una strategia interpretativa, la sua assenza (o presenza) nei confronti dei robot non implica nulla riguardo ai processi cerebrali o cognitivi: si apre, quindi, la possibilità di trattare l'assenza di atteggiamento intenzionale nelle descrizioni linguistiche del comportamento del robot come un fenomeno indipendente, dal momento che l'*intentional stance* non è una proprietà intrinseca degli agenti, né un meccanismo mentale radicato in specifici circuiti neurali, ma una strategia *astratta* adottata dall'osservatore nutrita dalle configurazioni concettuali della *folk psychology*. La criticità a cui è più difficile sopprassedere in Marchesi et al. (2021a), che non è presente in Marchesi et al. (2019), riguarda il legame diretto e causale che viene istituito tra misure implicite e presenza dell'atteggiamento intenzionale; se, d'altra parte, rileggiamo l'esperimento alla luce delle precisazioni sul concetto che abbiamo trovato in Dennett (1987), allora la forma implicita di misurazione non dice niente rispetto alla presunta adozione dell'atteggiamento stesso. D'altra parte, se possiamo provare a ricalibrare l'interpretazione modulando le istanze teoriche in gioco, non si possono ignorare i risultati empirici emersi da questi esperimenti: c'è una contraddizione, come riportato non solo da Marchesi et al. (2021a), ma anche altrove (Haring et al. 2013; Rosenthal-von der Pütten et al. 2013; Bartneck et al. 2009; Spatola et al. 2018, 2021b; Wykowska et al. 2014, 2015, 2016; Sciutti et al. 2018), in merito alle risposte esplicite ai test e quelle implicite. In sostanza, nella maggior parte degli studi sembra delinearsi una situazione in cui alle misure implicite che registrano le reazioni fisiologiche dei soggetti corrispondono risultati positivi circa l'attribuzione di intenzionalità, mentre allo stesso tempo le misure esplicite basate su questionari e test linguistici registrano un rifiuto a adottare la stessa posizione e tendono a negare il proprio coinvolgimento intenzionale nell'interazione con il robot. Si registrano, tuttavia, anche risultati opposti: alcune evidenze (Eyssel & Kuchenbrandt 2012; Severson & Carlson 2010; Sciutti et al. 2021) testimoniano di circostanze in cui i soggetti mostrano una tendenza all'uso di un linguaggio antropomorfo nei confronti del robot ma misure implicite come la registrazione dei tempi di risposta o il tracciamento dello sguardo non supportano la tesi della percezione intenzionale. Questo aspetto sottolinea una complessità di fondo del fenomeno, che è dimostrato dipendere anche da numerosi fattori 'esterni' e dalla conformazione fisica del robot (cfr. Perez-Osorio & Wykowska 2020; Rosselli et al. 2023;

Broadbent 2017). Ciò che si prospetta, allora, è la necessità di trovare una formulazione di un concetto di intenzionalità che renda conto della contraddittorietà delle prove empiriche nel doppio livello implicito ed esplicito, e che auspicabilmente riesca a rendere conto dei motivi per cui questo accade.

In ultima analisi, in aggiunta alla serie di riflessioni che abbiamo elaborato in merito agli esperimenti realizzati dal gruppo di Genova, questi studi ci consentono di vedere una potenzialità nell'ibridazione di diversi approcci all'intenzionalità, suggerendoci un'intuizione pratica: sebbene abbiamo dimostrato che ibridare l'approccio neo-cartesiano con quello neo-comportamentista sull'intenzionalità appare inconciliabile nei termini della teoria, forse altri tipi di ibridazioni sono possibili, e addirittura utili, per spiegare i risultati spesso incongruenti che si riscontrano negli esperimenti circa l'attribuzione di intenzionalità ai robot sociali.

In sintesi, la situazione che si delinea richiede di affrontare la situazione – epistemologica e sperimentale – riguardo la natura e il ruolo dell'intenzionalità nell'HRI perseguendo due obiettivi:

- i) Il primo obiettivo prevede di fornire una spiegazione dell'*intentional stance* che non sia naturalizzata in termini di processi cerebrali interni o che faccia appello a una qualche forma di linguaggio del pensiero interno alla mente, ma piuttosto che ne preservi la natura esterna come strategia linguistica alimentata dalla psicologia del senso comune.
- ii) Il secondo obiettivo richiede di gestire una situazione sperimentale in cui emergono risultati contraddittori a livello personale e subpersonale riguardo la comprensione del robot come agente intenzionale.

Tenteremo, quindi, nei prossimi paragrafi, di costruire una nuova ibridazione, che mantenga i presupposti dell'approccio neo-comportamentista all'intenzionalità ma che si apra ad accogliere i principi enattivisti della cognizione, che all'intenzionalità ha pensato come una proprietà prima di tutto corporea, frutto del nostro essere immersi nel mondo e continuamente presi in uno scambio interazionale con l'ambiente e con gli organismi che, insieme a noi, lo abitano. In questo modo, cercheremo anche di seguire il consiglio di Dennett per provare mostrare quello che, solitamente, nell'indagine maggioritaria sui modelli intenzionali, rimane invisibile.

### **4.3 Dall'attribuzione all'individuazione: un approccio enattivista per comprendere la HRI**

#### *4.3.1 L'intenzionalità enattivista*

Nelle scienze cognitive, l'approccio enattivista alla cognizione ha sviluppato considerazioni in merito all'intenzionalità fin dalle prime riflessioni in merito alla natura della mente. Già nel testo

fondativo dell'approccio, infatti, Francisco Varela, Eleanor Rosch ed Evan Thompson (1991) definiscono l'intenzionalità in aperto contrasto con le teorie classiche delle scienze cognitive, che allora costituivano l'approccio maggiore sulla totalità degli argomenti riguardanti la mente. A partire da una profonda ristrutturazione del panorama concettuale del tempo, e proprio in corrispondenza con una radicale riformulazione del rapporto di *aboutness* tra mente e mondo, i fondatori della corrente nelle scienze cognitive definiscono il loro approccio secondo cui la cognizione è *enazione*, ovvero una storia di accoppiamenti strutturali tra organismo e ambiente che fa emergere il mondo del significato, che si realizza tramite un network dinamico di sottosistemi sensorimotori interconnessi da cui emerge la capacità cognitiva dell'organismo (1991: 206). Nelle parole degli autori,

In a nutshell, the enactive approach consists of two points: (1) perception consists in perceptually guided action and (2) cognitive structures emerge from the recurrent sensorimotor patterns that enable action to be perceptually guided. (Varela et al. 1991: 173).

Nella loro proposta, rispetto a una visione che concepiva l'intenzionalità come una proprietà esclusiva degli stati mentali e in cui questa veniva spesso identificata con la capacità di rappresentare il modo al fine di costruire simboli da computare internamente, l'enattivismo apre a una revisione del concetto, riconfigurandolo all'interno della nuova prospettiva che vede la cognizione prima di tutto come un'*azione incarnata*, in cui i processi percettivi e cognitivi non si danno disgiunti o consequenzialmente, ma all'interno di un loop sensorimotorio garantito dall'*embodiment* e dalla *situatedness* dell'organismo nel proprio ambiente:

We would say that the intentionality of cognition as embodied action consists primarily in the *directedness of action*. Here the two-sidedness of intentionality corresponds to what the system takes its possibilities for action to be and to how the resulting situations fulfill or fail to fulfill these possibilities. (Varela et al. 1991: 206, corsivi nostri)

L'intenzionalità è qui, quindi, considerata come una capacità, un *know-how* messo in pratica dall'organismo nelle azioni che realizza all'interno dell'ambiente, identificabile non più come una proprietà statica della mente racchiusa nella scatola cranica o riducibile ai suoi stati mentali, ma come un processo dinamico in evoluzione e realizzantesi nell'emersione di ciò che è significativo per l'organismo.

Come per altri concetti propri della tradizione enattivista che trovano la loro definizione in questo testo, soprattutto in merito i) al significato da attribuire al termine *embodied*, nel suo doppio senso di incarnazione esperienziale di un corpo vissuto e corpo come elemento dinamico nel contesto dei meccanismi cognitivi (1991: xvi); ii) alla riconfigurazione del processo percettivo, secondo cui «perception is not simply embedded within and constrained by the surrounding world; it also contributes to the enactment of this surrounding world» (1991: 174); e iii) al rapporto tra percezione

e cognizione<sup>144</sup>, intimamente connessi nell'emersione dell'esperienza vissuta, anche quello di intenzionalità deve molto alle riflessioni fenomenologiche di Merleau-Ponty. In particolare, a caratterizzare la concezione enattivista di intenzionalità è il concetto di *intenzionalità motoria*, elaborato da Merleau-Ponty (1945) in opposizione a quello di intenzionalità cognitiva: mentre la seconda denoterebbe forme di intenzionalità più elaborate, che necessitano del linguaggio per essere espresse, la prima individua un tipo di intenzionalità corporea, incarnata, la cui attività può essere descritta come quella di costruire «un mondo, cioè un insieme di cose che emergono dall'informe proponendosi al nostro corpo come cose “da toccare”, “da prendere”, “da valicare”» (1945: 527-528), o anche quella modalità impiegata dal corpo per trovare il *grip* ottimale nell'ambiente (cfr. Dreyfus 2007; Rietveld 2012: 215-218). In sostanza, accanto a forme di intenzionalità linguistiche, si posiziona una forma di intenzionalità corporea che si basa sull'*engagement* dell'organismo con l'ambiente e che si esprime nelle azioni che l'organismo compie.

Sulla base di questo concetto, numerosi sono stati gli sviluppi che hanno portato alla costruzione di uno specifico approccio enattivista all'intenzionalità (Gallagher 2017, 2021; Hutto 2008; Hutto & Myin 2013, 2017; Steiner 2019, 2023, 2025). Possiamo tentare di riassumere i suoi principi fondamentali attraverso tre punti:

- i) *L'intenzionalità enattivista è una proprietà dell'intero organismo, non riducibile a processi fisici o funzionali nel cervello.*

A differenza di quanto espresso dall'approccio neo-cartesiano all'intenzionalità, per l'enattivismo questa è una caratteristica espressa da tutto l'organismo nella rete di *engagement* che si allaccia tra corpo, cervello e mondo e che dà come *uno* dei risultati la capacità cognitiva di dare senso all'ambiente all'interno del quale l'organismo è inserito: il *sense-making* (Thompson & Stapleton 2009; Di Paolo & Thompson 2014), cioè «the enaction of meaningful world by an autonomous agent» (Froese & Di Paolo 2011: 7). Tale caratterizzazione è già presente nelle prime formulazioni date da Varela, il quale definisce l'intenzionalità come una proprietà inerente al *sense-making*, che emerge nel processo di distinzione tra mondo e ambiente sulla base della capacità *embodied* dell'organismo: «The difference between environment and world is the surplus of signification which haunts the understanding of the living and of cognition, and which is at the root of how a self becomes one. In other words, this surplus is the mother of intentionality» (Varela 1992a: 7).

---

<sup>144</sup> «By using the term *embodied* we mean to highlight two points: first, that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context. By using the term *action* we mean to emphasize one again that sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition. Indeed, the two are not merely contingently linked in individuals; they also have evolved together» (Varela et al. 1991: 173).

In un approccio che vede la cognizione come emergente da e radicata nell'interazione dinamica dell'organismo autopoietico ed autonomo con l'ambiente in cui è inserito, l'intenzionalità non può che caratterizzarsi anch'essa come qualità scaturita da questo intreccio tra l'agente, la nicchia ecologica e i loop di percezione e azione enattivati al suo interno, esprimendosi in un processo di co-costruzione di significati. L'intenzionalità non è proprietà localizzabile nel cervello, negli stati mentali o nelle funzioni che questi espletano quando logicamente relati, ma una *direzione di senso* che si sviluppa nella relazione concreta con l'ambiente:

[C]oupling is possible only if the encounters are embraced from the perspective of the system itself. This amounts, quite specifically, to elaborating a surplus signification relative to this perspective. Whatever is encountered must be valued one way or another—like, dislike, ignore—and acted on some way or another—attraction, rejection, neutrality. This basic assessment is inseparable from the way in which the coupling event encounters a functioning perceptuo-motor unit, and it gives rise to an *intention* (I am tempted to say “desire”), that unique quality of living cognition (Dennett 1987). (Varela 1992a: 12)

Attraverso l'impiego di una diversa terminologia, la stessa idea è portata avanti da Gallagher, il quale argomenta anche in merito alla caratteristica che sembra definire, almeno negli approcci *mainstream*, l'intenzionalità come unico e solo marchio del mentale. Al contrario, Gallagher sottolinea che, nell'approccio enattivista, oltre a non essere riducibile a pattern neuronali o funzionali, la caratteristica principale dell'intenzionalità è quella di essere una proprietà relazionale della cognizione, ed è solamente insieme a tutte le altre proprietà emergenti dal network creato nell'interazione di corpo, cervello e mondo che può identificarsi la mente:

[...] non-derived intentionality is not reducible to either physical or functional processes in the brain, but that it is better understood as a more embodied form of transaction with the environment. Non-derived intentionality, as a basic motor intentionality, *is* the action-oriented coupling of organism-environment. Furthermore, even on this notion of intentionality, neither non-derived nor derived intentionality is the sole mark of the mental. [...] intentionality is only one possible element of a pattern that constitutes the mind. (Gallagher 2017: 64)

ii) *L'intenzionalità enattivista si esprime in interazione con l'ambiente e con gli altri tramite le azioni realizzate dell'agente.*

Gallagher (2017, 2021) è senza dubbio colui che ha più sviluppato questa visione sull'intenzionalità, costruendo un approccio enattivista alla cognizione sociale basato sulla critica agli approcci folkpsicologici e simulazionisti e abbracciando una visione della comprensione intersoggettiva basata innanzitutto sulla *percezione diretta* delle intenzioni dell'altro nei movimenti realizzati durante l'interazione, ovvero una percezione senza necessità di inferenza che la renda significativa (2021: 131). L'idea è esattamente che «the meaning, the intentionality of an action, is perceptible» (2021: 123) dal momento che

a) L'intenzionalità crea la cinematica dell'azione, veicolando già l'obiettivo dell'azione stessa nel momento in cui viene svolta secondo alcune modalità (e non altre; Becchio et al. 2012, 2017), come dimostrato da diversi studi che evidenziano la capacità degli individui di identificare un movimento intenzionale tramite la specificazione di dinamiche cinetiche differenti (Ansuini et al. 2008; Sartori et al. 2011). È in questo senso che si può intendere la frase, diventata un motto per gli approcci della percezione diretta alla cognizione sociale, che «intentions become 'visible' in the surface flow of agents' motions» (Ansuini et al. 2014), accolto anche da Gallagher (2021: 124).

b) L'azione, negli scenari di interazione quotidiana con altri individui, non è mai svolta in scenari asettici e nudi, ma sempre in un contesto già pregno di significato, che a sua volta supporta lo svolgimento di alcuni tipi di azioni (e non altri). Inoltre, nella maggior parte dei casi, le azioni non sono svolte a seguito di una pianificazione deliberata, e a volte nemmeno in maniera propriamente razionale: «In everyday circumstances we are not dealing with the kind of abstractions that we employ in experiments [...]. Most often actions are fully contextualized by physical, social, and cultural affordances» (2021: 125). A differenza degli approcci internalisti, l'intenzionalità qui diventa comprensibile in interazione, proprio perché l'incontro reciproco e situato nell'ambiente costituisce uno degli elementi costitutivi della cognizione sociale.

Nella formulazione di Gallagher, e nella sua *Interaction Theory* (*infra* §3.3), la nozione di intenzionalità sviluppata si nutre fortemente delle riflessioni fenomenologiche di Merleau-Ponty, soprattutto in relazione alla nozione di intercorporeità. In particolare, questa viene definita come una reciproca capacità di comprendersi che emerge dall'incontro di due (o più) corpi *vivi*, in cui l'esperienza dell'altro si dà *con* l'altro:

Intercorporeity involves a mutual influence of body schemas, but not in isomorphic format, where one mirrors, or maps the other's actions onto one's own motor representations. Rather, intercorporeity involves a reciprocal dynamic, and enactive response to the other's action, taking that action as an affordance for further action rather than as the occasion for replication (simulation). (Gallagher 2017: 77).

Quello che facciamo quando cerchiamo di capire qualcuno in interazione, dunque, non consiste in un processo in cui, in primo luogo, percepiamo movimenti non intenzionali da parte del sistema e, di seguito, cerchiamo di trovare una interpretazione tramite inferenze circa la causa, l'intenzione o il motivo di quelle azioni; percepiamo enattivamente quei movimenti come forme di intenzionalità, ovvero immediatamente significative poiché inserite all'interno di un'ecologia relazionale, interazionale, materiale, sociale e culturale che ne condiziona le possibilità di significare. Ancora seguendo Gallagher, «Enactive perception of others means that we see their

emotional expressions and contextualized actions as meaningful in terms of how we might respond to or interact with them» (Gallagher 2017: 77-78).

In altre parole, nel contesto della cognizione sociale, questo significa che il modo in cui principalmente comprendiamo gli altri in quanto agenti intenzionali è tramite un concetto di *intenzionalità operativa* (Gallagher 2017: 76), che è intrinseca non agli stati mentali ma al movimento, non basata sull'attribuzione di stati mentali ma percepita nei termini della sua adeguatezza al contesto interazionale, al suo valore pragmatico, condizionato dalla situazione e dalla normatività che viene importata nell'interazione oltre che diretto al compimento di un'azione rivolta nei confronti dell'altro o dell'ambiente che gli agenti co-abitano<sup>145</sup>.

Intentional bodily movements [...] have very distinctive properties that reflect intention; they are simultaneously constrained by the agent's goal, by the attributes of the situation, and by a set of kinematic and biochemical rules that jointly shape their dynamics [...]. The intentional aspects of bodily movements are not extrinsic to those movements – they are intrinsic and are reflected in their organization. Intentional actions have observable characteristics that distinguish them from non-intentional behaviors. (Gallagher 2021: 127)

Nel quadro della percezione dell'intenzionalità dell'altro, un'ultima precisazione va rivolta alla *primarietà* che in questo punto di vista viene riservata al tipo di intenzionalità. Infatti, questa visione corporea e interattiva, che rende l'intenzionalità operativa, è considerata primaria (Gallagher 2017, 2021) in un duplice senso: non solo è il solo modo in cui comprendiamo il mondo e gli altri durante l'intersoggettività primaria (Trevarthen 1979), ma anche in età adulta rappresenta la prima modalità a cui facciamo affidamento per comprendere il mondo e gli altri, ovvero quella corporea e interattiva. Questo è collegato a un cambio di prospettiva negli studi sulla cognizione sociale, che abbiamo già sottolineato nel capitolo precedente (cfr. *infra* §3.3.4): il movimento che si produce passa dalla visione in terza persona degli altri approcci esaminati finora, in cui l'osservazione è analizzata dall'esterno tramite un occhio terzo che legge l'incontro mentre non è coinvolto, alla prospettiva in prima e seconda persona introdotta nell'enattivismo, soprattutto attraverso le concezioni fenomenologiche dell'intersoggettività che vede gli effetti dell'embricazione del soggetto nell'incontro con l'altro prima di tutto nella strutturazione dell'esperienza individuale.

---

<sup>145</sup> «Actions have intentionality because they are directed at some goal or project, and this is something that we can see in the actions of others. With respect to understanding the specific claim that in contexts of interaction one can directly perceive the intentions of others (without needing to infer or simulate them), the idea is that the meaning, the intentionality of action is perceptible. To claim that the meaning of something is perceptible requires that we define perception as rich, enactive, or as I'll say, "smart" process, and not just a registration of sensory input». (Gallagher 2021: 123)

- iii) *L'intenzionalità enattivista recupera un diverso significato di aboutness, che non implica necessariamente l'impiego di rappresentazioni mentali.*

Questo aspetto è stato approfondito da Hutto e Myin (2013, 2017; Hutto 2008; Hutto & Satne 2015), secondo cui «it is conceptually possible to disentangle ideas that unhelpfully run together in Brentano's classical formulation of the various properties of intentionality» (Hutto & Myin 2017: 94), che hanno condotto a un legame quasi inscindibile tra intenzionalità e rappresentazionalismo nelle scienze cognitive classiche. Quella che propongono gli autori, invece, è una intenzionalità non-rappresentazionale, in linea con i principi corporei e situati descritti fino a questo momento, ma radicalizzando la necessità di espellere completamente il contenuto proposizionale da queste forme basilari di cognizione: in sostanza, una *contentless intentionality*, definita *Ur-intentionality* (Hutto 2008; Hutto & Myin 2017: 93-120; Hutto & Satne 2015). Ciò è reso possibile grazie al fatto che forme basilari di intenzionalità non hanno, secondo gli autori, la necessità di essere definite secondo le loro condizioni di correttezza o di soddisfazione, ma vanno intese come configurazioni di significato realizzate in modalità teleosemiotiche di accoppiamento con l'ambiente (cfr. *infra* §4.3.2). Nelle parole di Hutto e Myin (2017: 101), «this is more or less the REC view: in basic kinds of cognition an organism's skillful engagements with the world are best understood in embodied, enactive, and nonrepresentational ways».

L'*aboutness*, ovvero la caratteristica del pensiero di essere necessariamente riguardo qualcosa, non deve essere più specificata in termini di contenuti proposizionali, stati mentali o condizioni di soddisfazione di verità, ma diventa la direzionalità semiotica dell'azione scaturita da una spinta biologica al *sensemaking*. In questo processo, risulta indispensabile la nozione di *affordance*, cioè «the possibilities for action provided to us by the environment» (Rietveld et al. 2018: 42), che rende intelligibile il significato in quanto azione realizzata nel mondo come qualcosa che non ha sede specifica: «The so-called meaning generated through sense making is thus neither a feature of the external environment nor something internal to the agent» (Hutto & Myin 2017: 78). Il modo in cui gli organismi sono situati e incarnati nell'ambiente seleziona una congerie di possibilità di azione (*affordance*) che vengono percepite dagli stessi organismi come eseguibili all'interno del proprio ambiente; allo stesso tempo, su una scala temporale evolutiva, l'ambiente seleziona le caratteristiche degli organismi più congeniali e ne dirige la adattabilità, in un processo di *natural drift* di co-aggiustamenti sub-ottimali che si realizzano dinamicamente (Varela et al. 1991: 185-216).

L'intenzionalità, in questo senso, restituisce un significato diverso al concetto di *aboutness* del pensiero, originariamente formulato da Brentano. In questa prospettiva, le rappresentazioni non sono uno strumento concettuale indispensabile per collegare la mente all'oggetto intenzionato,



poiché la mente – non confinata all'interno della pelle e del cranio dell'individuo – è sempre immersa nel mondo, e insieme ad esso concorre a creare le condizioni per la cognizione. È solo comprendendo la mente in questa prospettiva che è possibile individuare la natura 'correlazionale' dell'intenzionalità (Thompson 2007: 26; Dupuy 1994), che non si dà necessariamente come qualità intrinseca alla mente e supportata da rappresentazioni simboliche degli eventi esterni, ma come relazione tra la capacità del sistema di dare senso e le occasioni di senso raccolte nell'ambiente.

[The meaning of external events] or significance corresponds to an attractor of the system's dynamics (a re current pattern of activity toward which the system tends), which itself is an emergent product of that very dynamics. The external world is constituted as such for the system by virtue of the system's self organizing activity. Dupuy's proposal, in a nutshell, is that constitutional intentionality corresponds to a kind of self-organization. (Thompson 2007: 27)

Riassumendo, queste sono le tre principali caratteristiche dell'intenzionalità enattivista, che si differenzia rispetto a tutti gli altri approcci delineati fin qui soprattutto per la sua natura relazionale e dinamica. Infatti, mentre l'approccio neo-cartesiano che si caratterizza come statico e in cui l'intenzionalità è resa necessaria come strumento concettuale che istituisce la comunicazione tra mondo esterno e mente interna all'individuo, l'intenzionalità enattivista supera il dualismo mente-mondo e offre una visione della mente come intrinsecamente legata al corpo e all'ambiente in cui è immersa. In questo quadro, l'intenzionalità non si basa su rappresentazioni interne ma su un accoppiamento attivo e dinamico tra organismo e ambiente, fattore che risulta fondamentale per spiegare fenomeni cognitivi complessi come l'apprendimento motorio, la percezione e l'adattamento in tempo reale. Inoltre, se il neo-cartesianesimo si fondava sulle rappresentazioni simboliche come elementi costitutivi dell'intenzionalità, l'approccio enattivista ne dissolve la necessità, spostando l'attenzione dal *come* le rappresentazioni mentali corrispondono al mondo, al *come* l'organismo interagisce con il mondo per produrre significati attraverso l'azione incarnata e situata.

D'altra parte, rispetto all'approccio neo-comportamentista basato sull'attribuzione, l'intenzionalità così definita rende possibile la comprensione di interazioni complesse che vanno oltre i meccanismi interpretativi basati sull'osservazione dello svolgimento del comportamento. Per comprendere questa differenza, si prenda ad esempio la circostanza in cui mi trovi a parlare della mia tesi di dottorato ad un gruppo di amici che non conosce la filosofia della mente o l'HRI e mi accorgessi, mentre parlo, che li sto annoiando a morte. Decido quindi di cambiare discorso, iniziando a parlare di tematiche che so che possono essere più interessanti per tutti loro. Questa decisione richiede una serie di processi cognitivi complessi: devo leggere i segnali corporei dei miei amici (ad es. l'espressione facciale, il linguaggio del corpo, la direzione dello sguardo, ecc.), confrontare questi segnali con la mia esperienza precedente (che mi permette, ad esempio, di riconoscere quando

qualcuno è annoiato o poco interessato alle mie parole) e poi formulare una nuova risposta o pensare a una diversa direzione per la conversazione. Dal punto di vista neo-comportamentista, si potrebbe descrivere la situazione così: adottando un atteggiamento intenzionale nei confronti dei miei amici, sulla base del comportamento che osservo attribuisco loro su base induttiva una serie di stati mentali (come il disinteresse, la noia) e scommetto che, cambiando il mio comportamento, anche il loro mostrerà segni di un più interessato coinvolgimento. Uso, quindi, l'inferenza ai loro stati mentali come credenze ("questa conversazione è molto noiosa, non mi importa niente delle conseguenze interazionali dei modelli epistemologici impiegati") e desideri ("vorrei moltissimo che Martina la smetta di parlare di interazioni tra umani e robot) al fine di costruire la mia strategia di compensazione e non farmi odiare dai miei amici. La situazione descritta, come è chiaro, si poggia su una estrema razionalità dell'agente che adotta l'atteggiamento intenzionale, il quale dalla descrizione del processo sembra aver elaborato con cura la propria strategia e aver ragionato attentamente sugli scenari futuri legati alle sue azioni possibili. Tuttavia, sappiamo bene che queste circostanze sono molto rare: nelle interazioni quotidiane, e anche in quelle che prevedono una buona dose di previsioni e ragionamento, il modo in cui comprendiamo gli altri e ci aggiustiamo in corrispondenza dei loro atteggiamenti non è basata su processi inferenziali. Il vantaggio dell'approccio enattivista all'intenzionalità consiste esattamente nell'introdurre la possibilità di descrivere, spiegare e comprendere situazioni come quella appena mostrata in termini di processi dinamici di interazione. In questa specifica situazione, la mia decisione di cambiare argomento emerge dall'accoppiamento strutturale tra me e i miei amici, basato su una continua interazione incarnata e sul contesto che accoglie la nostra interazione. Non sono obbligata ad attribuire specifici stati mentali ai miei amici, ma reagisco in modo fluido e immediato ai segnali corporei e contestuali che insieme co-creiamo. L'intenzionalità è quindi qualcosa che emerge dall'interazione stessa, attiva e incarnata, e viene distribuita al suo interno, non qualcosa che anche nelle forme più corporee deve essere attribuita tramite inferenza.

Infine, l'approccio neo-pragmatista all'intenzionalità, ponendo l'attenzione sulle pratiche sociali e linguistiche, ben rende conto della matrice sociale del significato, ma trascura la componente incarnata della cognizione. L'intenzionalità enattivista, d'altra parte, integrando la dimensione corporea e quella sociale (Froese & Di Paolo 2011; Di Paolo et al. 2007; 2018), evidenzia come l'accoppiamento sensorimotorio con l'ambiente e con gli altri sia un aspetto cruciale nel processo di costruzione del significato. Si potrebbe argomentare, come fatto altrove, che l'approccio enattivista amplii quello neo-pragmatista, fornendo una base corporea alla creazione di significati socialmente e culturalmente condivisi (Gallagher & Miyahara 2012; Gallagher 2017; Gallagher & Hutto 2008; Hutto 2008, 2009).

Ora, le implicazioni relative all'introduzione di una nuova concezione di intenzionalità non si esauriscono sul piano teorico. Già nella riflessione che inaugura l'alternativa enattivista, gli autori intravedono una serie di cambiamenti non solo per la riconsiderazione dell'intenzionalità come strumento epistemologico, ma anche nell'indagine generale dei processi cognitivi per come debbano essere affrontati da chi si occupa di comprendere i legami tra manifestazioni diverse della mente, ovvero percezione, cognizione e azione:

What does this reconceptualization of the intentionality of cognition imply in more pragmatic terms for cognitive science? Consider that there are two domains in which we can describe any cognitive system: on the one hand, we can focus on the *behavioral interactions* of the system by describing it as composed of various subsystems, etc., and on the other hand, we can focus on the behavioral interactions of the system by describing it as a unity capable of various forms of coupling. In switching back and forth between these two kinds of description, we – that is, cognitive scientists – must determine both how the environment constraints the system and how these constraints themselves are specified by the sensorimotor structure of the system [...]. In doing so, we are able to explain how regularities – sensorimotor and environmental – emerge from structural coupling. The research task in cognitive science is to make transparent the mechanisms by which such coupling actually unfolds and thereby how specific regularities arise. (Varela et al. 1991: 206)

Possiamo provare a tradurre la sfida lanciata oramai più di trenta anni fa dagli autori di *The Embodied Mind* negli studi sull'interazione tra umani e robot. Cosa comporta questa riconcettualizzazione dell'intenzionalità nel campo di studi che cerca di capire se trattiamo anche i robot, oltre che gli altri esseri umani, come agenti intenzionali? Il passaggio che abbiamo riportato illumina una doppia possibilità: da una parte, potrebbe condurci a comprendere come l'ambiente, che include gli organismi che lo abitano, vincola il sistema; dall'altra, ci permetterebbe di cogliere le modalità in cui i vincoli stessi sono specificati dalla struttura sensorimotoria del sistema cognitivo. In termini interazionali, questo significa porre l'attenzione sugli aggiustamenti che l'essere umano apporta in interazione con il robot, focalizzandoci sul modo in cui vengono sviluppate modalità di risposta e di aggiustamento al flusso dell'interazione. Così, l'enattivazione di determinate azioni di fronte al robot, registrabili a livello corporeo, potrebbero costituire una evidenza del modo in cui comprendiamo questi particolari agenti in interazione.

Con l'obiettivo di rendere conto di questo nuovo sguardo all'intenzionalità negli studi di HRI, ci rivolgeremo a una specifica formulazione del concetto di intenzionalità sviluppata negli ultimi anni all'interno di quella che si auto-proclama la corrente dell'enattivismo radicale (*Radical Enactive (Embodied) Cognition*, REC), la quale sostiene che

A truly radical enactivism – REC – holds that it is possible to explain a creature's capacity to perceive, keep track of, and act appropriately with respect to some object or property without positing internal structures that function to represent, refer to, or stand for the object or property in question. Our basic ways of responding to worldly offerings are not semantically contentful. Still, it is possible to explain how perception gets its work

done in guiding cognitive activity – even when it is quite sophisticated – without assuming the truth of CIC [Content Involving Cognition]. (Hutto & Myin 2013: 82)

Analizzando come la REC ha declinato la visione enattivista dell'intenzionalità, nei prossimi paragrafi ci concentreremo su due aspetti fondamentali. Da un lato, delineeremo un cambiamento nel modo in cui concepiamo il processo di comprensione del robot in quanto sistema e agente intenzionale, inclusa la terminologia impiegata per descrivere il fenomeno in questione (§4.3.2). Dall'altro lato, partendo da questo cambiamento, scopriremo come sia possibile riconsiderare il ruolo dell'intenzionalità e le modalità di indagine sul fenomeno del suo rivelamento negli esperimenti con robot sociali, aprendo così a nuove opportunità di comprensione (§4.3.3).

#### 4.3.2 *Intentionality Detection: forme di riconoscimento dell'intenzionalità*

Per comprendere al meglio come la REC declini il concetto di intenzionalità, bisogna innanzitutto rendere chiari gli assunti su cui questo approccio alla cognizione si regge. In prima istanza, la radicalizzazione dello sguardo è proposta al fine di liberare l'enattivismo da qualsiasi residuo di computazionalismo, rappresentazionalismo o riferimento a correlati funzionali rintracciabili nel cervello, almeno per quello che riguarda la struttura e la composizione delle forme più incarnate di cognizione.

Since certain prominent variants of enactivism harbor conservative tendencies and thus fail to break with tradition in the required way, it is necessary to press for a radicalized enactivism [...], which denies that basic minds are contentful. (Hutto & Myin 2013: xii)

Il discorso, come è possibile notare, ruota intorno alla nozione di *basic mind*. Questa, in particolare, presenta diverse caratteristiche: i) non implica o include alcun tipo di contenuto<sup>146</sup> (2013: 1); ii) è costitutivamente e fondamentalmente *world-involving*, ovvero si realizza nel coinvolgimento dell'organismo con l'ambiente e non è localizzabile in porzioni definite del cervello o degli stati funzionali dell'organismo (2013: 137); iii) è non-rappresentazionale, dal momento che le rappresentazioni sono dimostrate essere inessenziali per la formazione e l'indirizzamento delle azioni e dell'esperienza dell'organismo (2013: x; 2017: 39); iv) è basata sugli abiti d'azione e di interpretazione: «RECCers assume that many sophisticated embodied engagements constitute mentality and are best explained with reference to nothing more than habits of mind, without the need to invoke the existence of any content-involving or representationally informed cognition whatsoever» (2013: 15).

---

<sup>146</sup> Il contenuto, in linea con la tradizione analitica della filosofia della mente, è definito nella relazione con le condizioni di verità: «At its simplest, there is content wherever there are specified conditions of satisfaction» (Hutto & Myin 2013: x).

In altre parole, le *basic mind* sono considerate la forma più essenziale di cognizione, comune (almeno) alle forme di vita animali, responsabile delle forme basilari di comprensione e intervento sull'ambiente. La *basic cognition* si realizza nei coinvolgimenti sensibili e selettivi in un processo di costruzione del significato incarnato, è un processo attivo di configurazioni *world-involving*, è costitutivamente interattiva, dinamica e relazionale e è al cuore di forme di cognizione umana che includono la percezione, l'immaginazione e la memoria, sia nei momenti dello sviluppo sia in età adulta.

Nel quadro delle *basic mind*, l'intenzionalità si declina proprio come una qualità fondamentale della *basic cognition*, la quale esibisce una *intentional directedness* senza implicare rappresentazioni o contenuto.

Fundamentally, by REC's lights, basic cognition is a matter of sensitively and selectively responding to information, but it does not involve picking up and processing information or the formation of representational contents. REC's account of basic cognition is thus given in terms of active, informationally sensitive, world-involving engagements, where a creature's current tendencies toward active engagement are shaped by its ontogenetic and phylogenetic history. (Hutto & Myin 2017: 92)

La capacità delle *basic mind* di rivolgersi a oggetti e obiettivi nell'ambiente e di essere caratterizzate da un'intenzionalità *world-involving* è garantita dalle nozioni di i) *teleosemiotica* e di ii) *Ur-intenzionalità*. Per quanto riguarda a), il concetto è stato sviluppato in contrasto alla teoria teleosemantica di Millikan (2005), che ha lo scopo di fornire spiegazioni riguardo a come le rappresentazioni mentali e linguistiche acquisiscano contenuti oggettivi, ovvero come gli stati mentali possano arrivare a rappresentare oggetti del mondo<sup>147</sup>. Il suo obiettivo principale è quello di fornire una spiegazione naturalizzata del contenuto mentale e del significato, giustificandola all'interno di una concezione rappresentazionalista e computazionalista biologicamente radicata della mente. Al contrario, dunque, per la REC «there is not apparent scientific reason for supposing that the connections between an organism and its world that are established by Mother Nature are best characterized in semantic terms, such as reference or truth» (Hutto & Myin 2017: 111). La teleosemiotica prende quindi il posto della teleosemantica come teoria atta a spiegare l'intenzionalità e l'emersione del significato non in termini rappresentazionali o verocondizionali: per la teleosemiotica, il contenuto non è determinato da una funzione rappresentazionale evolutivamente

---

<sup>147</sup> Le teorie teleosemantiche del contenuto stabiliscono un rapporto diretto tra funzioni biologiche e rappresentazioni mentali: «Construed in semantic terms, such theories conceive of mental representations as inner states that have the biological function of enabling organisms to keep track of specific worldly items. Proponents of traditional teleosemantic theories understand the content of mental representations in truth condition terms – thus a mental representation is true if things are the way an inner state represents them as being, and it is false in all other cases» (Hutto & Myin 2017: 105). La neve è bianca se e solo se la neve è bianca.

selezionata, ma emerge dalle interazioni immediate e pratiche che un organismo ha con il proprio ambiente.

[Biological capacities] can be understood in biological terms as mechanisms through which basic minds are *set up to be set up* by other minds and *to be set off* by certain things. As a baseline, all that is required is that basic minds can target and respond to certain things in way relevant for learning various tasks from their fellows. According to REC's teleosemiotic view, individuals can respond to the environment and to each other in ways that allow for emulation, imitation, and regulation of what they target and attend to in ways that make basic forms of social learning possible. (Hutto & Myin 2017: 140)

Il significato non è quindi una proprietà intrinseca di un segnale, dell'informazione proveniente dall'ambiente o di una rappresentazione, ma si costituisce attraverso le dinamiche normative e contestuali che caratterizzano l'azione e l'interpretazione.

L'intenzionalità si caratterizza quindi come (ii) *Ur-intenzionalità*, ovvero una forma primitiva di intenzionalità che permette alle *basic mind* di attenzionare specifici oggetti e stati di cose nell'ambiente senza necessariamente rappresentarle a livello di contenuto, manifestandosi nelle capacità interattive dirette degli organismi. Questa intenzionalità di base è strettamente connessa all'agire nel mondo: l'organismo non rappresenta simbolicamente l'ambiente, ma si embrica con esso tramite azioni *embodied*, che riflettono la sua storia di interazione.

Questa forma di intenzionalità, dunque, deriva dalle relazioni dinamiche tra organismo e ambiente, dove il significato emerge *in situ*, senza dover fare riferimento a processi computazionali o a strutture semantiche predefinite. Ed è anche ciò che ci permette di comprendere le azioni degli altri senza far riferimento a ragionamenti inferenziali o pratiche di *mindreading* come invece ipotizzato nella ToM. A partire dall'analisi di come i bambini apprendono la capacità di comprendere gli altri nelle loro azioni complesse e di interpretare il loro comportamento tramite l'utilizzo del linguaggio, Hutto (2008) elabora una configurazione della cognizione basata su due capacità o caratteristiche fondamentali:

- *Intentional attitude*. A differenza della caratterizzazione proposta da Dennett (1971, 1987), per Hutto l'atteggiamento intenzionale non si esprime nelle capacità di elaborare strategie facendo ricorso a concetti folkpsicologici come gli stati mentali, ma è una capacità strettamente legata alla modalità di essere inseriti nel mondo dell'organismo animale e delle *basic mind* in generarle<sup>148</sup>, cioè una declinazione dell'Ur-intenzionalità nel campo intersoggettivo. Hutto le definisce come «a thoroughly noncognitivist, nonrepresentationalist understanding intentionality – one that regards embodied, enactive modes of responding as basic and sees

---

<sup>148</sup> «The difference is that REC conceives of the commonalities present in all forms of cognition in terms of interactive engagements (with the world, with others), as opposed to understanding such commonalities in terms of mental states with representational content. Thus, REC assumes that there is a basement-level similarity between all cognitive creatures, and that – *if we focus on those basic features of cognition* – human beings are not cognitively unique!». (Hutto 2008: 136)

symbolic thinking as the preserve of those beings that have appropriately mastered certain sophisticated linguistic constructions and practices» (Hutto 2008: xiii) e le connota di alcune caratteristiche identificative: le *intentional attitudes* sono dirette a possibili situazioni e configurazioni interazionali, non a proposizioni, ed esibiscono ‘naturalmente’ una direzionalità intenzionale (2008: 45). La manifestazione dell’*intentional attitude*, inoltre, è realizzata dall’organismo nel suo complesso, il quale mette in atto degli sforzi coordinati nella costruzione di comportamenti significativi:

The essential idea of my proposal is that both infants and adults are directly responsive to the psychological situation of others because they are informationally sensitive to a special class of natural signs – the expressions of intentional and affective attitudes, as revealed in another’s gaze, gesture, facial comportment, and so on. In one-to-one encounters of this sort participants embody intentional attitudes that are directed toward the intentional attitudes of others. To understand this proposal, it is important to recall that intentionality is not a property of functionally specified internal or inner “mental states”; only whole organisms and their coordinated efforts exhibit it. Intentionality and affect are therefore expressed by the way organisms carry themselves. (Hutto 2008: 117)

È questa la prima modalità, quella su cui facciamo costantemente affidamento non solo nel processo di sviluppo cognitivo che affrontiamo nelle prime fasi dell’intersoggettività (Trevvarthen 1979; Reddy 2008), ma anche nelle interazioni quotidiane in cui non concentriamo la nostra attenzione a voler scoprire le cause nascoste del comportamento dell’altro. È certo che, se voglio scoprire i motivi reconditi che spingono il mio amico Flavio a chiudere inesorabilmente lo sportello della lavatrice, dovrò fare appello a spiegazioni provenienti dal contesto socioculturale e linguistico a cui entrambi apparteniamo (ad esempio penserei: “vorrà farmi un dispetto?”, “avrà intenzione di farmi arrabbiare per vendicarsi di un torto?”, “mi vuole portare all’esasperazione?”), ma la maggior parte delle interazioni non ha bisogno di tanto sforzo cognitivo. Per capire le intenzioni di Flavio mentre mi sta parlando mi basta entrare in interazione con lui e individuare le sue *intentional attitudes*, attività connaturata alla mia stessa abilità cognitiva di animale umano, per capire se è triste, felice, se mi sta mentendo, se sta provando a farmi arrabbiare o se è interessato di più al piatto di pasta che si trova davanti a me. Nella chiarissima descrizione data da Hutto,

There is a simpler, and more attractive, way of understanding nonverbal cognition and its form of intentionality. Organisms of all kinds are informationally sensitive to select worldly offerings – sometimes in complex ways. These built-in activities permit them to identify and track certain environmental particulars of importance (to them); *detection* of these particulars initiates responses that enable the coordination of successful actions in certain environments. This package will have been selectively forged in the service of meeting the end-directed needs of specific creatures.” (Hutto 2008: 50; corsivi e grassetti miei).

Questa comprensione diretta è garantita dal fatto che gli organismi, esseri umani compresi, sono naturalmente sensibili a determinati aspetti del mondo che per loro hanno una qualsivoglia rilevanza, sancita nel corso della filogenesi dai processi evolutivi di co-aggiustamento con l'ambiente. Questa sensibilità consente loro di riconoscere e rispondere a determinati elementi ambientali, facilitando un'interazione efficace con l'ambiente e con gli altri organismi che lo co-abitano. L'intenzionalità che gli organismi mostrano, quindi, si esprime nel comportamento, e non è il risultato dell'intrecciarsi di stati mentali interni. Per esempio, quando vedo Flavio chiudere con forza lo sportello della lavatrice, la mia mente non deve analizzare internamente qualche motivazione astratta; piuttosto, la mia sensibilità percettiva alle sue *intentional attitudes* mi permetterà di interpretare il gesto come un atto compiuto in preda alla rabbia. La mia attenzione sarà catturata da questo gesto, forse inaspettato, perché, in quanto animali sociali, noi esseri umani siamo particolarmente sensibili alle modalità di espressione dei nostri consimili:

Human infants identify, attend, and respond to the intentional attitudes expressed in the goal-directed behavior of others in reliably expectant ways. They are naturally attuned and appropriately sensitive to the expressed attitudes of certain others, at least in certain types of circumstances. (2008: 101)

Infine, in linea con le prospettive generali dell'intenzionalità enattivista, anche le *intentional attitude* sono espresse dall'organismo nel suo complesso, manifestate in atteggiamenti comportamentali globali e visibili nel contesto di realizzazione:

Organisms are informationally sensitive *to* and selectively end-directed *at* certain worldly features, objects, and states of affairs in a way that explains their success and failure on certain tasks. The intentionality they exhibit is an attitude of the whole organism expressed in their behavior; it is neither a property of the signs themselves nor of organismic inner states. (2008: 57)

In sostanza e riassumendo, le *intentional attitude* sono quelle caratteristiche del comportamento espresse direttamente nel movimento dell'agente, che vengono riconosciute e attenzionate dall'altro poiché salienti in un sistema intersoggettivo alle cui dinamiche siamo abituati fin dai primi giorni di vita. Non hanno bisogno di linguaggio mentalistico per essere comprese, non si basano su un processo di attribuzione per essere rilevate: semplicemente, aderiscono al comportamento dell'agente in quanto intrinseche alla costruzione di senso. È così che l'altro in interazione può carpirle direttamente:

Well-calibrated social activity only involves a capacity to selectively respond to end-direct intentional attitudes that are revealed in the expressions of others; these expressive attitudes are unlike those of the propositional variety. For this reason, among others, these primary modes of interacting with others should not be classified as a species of mindreading. In "normal" contexts, such methods are highly effective – indeed, arguably the best – means of navigating social dynamics. (Hutto 2008: 3)



Ma come fare, allora, a rendere conto di forme più complesse di comprensione degli altri, che devono fare riferimento a norme sociolinguistiche e a un linguaggio specifico per essere espresse?

- *Propositional attitude*. Queste entrano in gioco proprio in corrispondenza di forme più complesse di comprensione, per cui non sono sufficienti o disponibili modalità dirette e corporee di intendimento. In particolare, le *propositional attitude* costituiscono un modo più evolutivamente sviluppato di comprendere e interpretare il comportamento degli altri, riferendosi a caratteristiche centrali di credenze e desideri. La capacità di gestire e sfruttare l'atteggiamento proposizionale viene acquisita nel processo di crescita dell'individuo, prendendo parte a quelle che Hutto definisce *narrative practice* (Gallagher & Hutto 2008: Hutto 2008; Hutto & Myin 2017). Le pratiche narrative, intese come «direct encounters with stories about persons who act for reasons» (Hutto 2008: x), rendono possibile l'accesso alla strutturazione basilare della *folk psychology* e alle opportunità accettabili di gestire tale struttura nelle pratiche, imparando quindi sia come sia quando utilizzare determinate porzioni di sapere organizzato in maniera proposizionale.

Long before they acquire a practical grasp of mentalistic concepts, children are able to navigate the social matrix using embodied skills, interacting with others in ways that require no understanding of propositional attitudes or reasons for action whatsoever. As their command of language increases, that are able to make use of certain syntactic constructions [...]. This brings new objects of attention and joint attention into view, allowing them to graduate from engaging with others exclusively by responding to intentional attitudes, to possessing the capacity to understand them in terms of propositional attitudes. (Hutto 2008: 141)

È attraverso l'incontro con questo tipo di storie, del tipo delle favole della buona notte raccontate dal *caregiver* in momenti di interazione attiva, che i bambini imparano a individuare le credenze e i desideri che spingono i protagonisti di queste storie a fare quello che fanno: «It is through scaffolded encounters with stories of the appropriate kind that children learn how the core propositional attitudes behave with respect to one another and other standard mental partners» (2008: xii). Le pratiche narrative forniscono un insieme di esempi indicativi per comprendere le ragioni degli altri, assumendo la funzione di modelli da cui verranno estratte le caratteristiche strutturali che agevoleranno la comprensione esplicita dei comportamenti tramite il dizionario della psicologia del senso comune; partecipando a queste storie, i bambini acquisiscono una comprensione diretta di come le attitudini proposizionali, incluse credenze e desideri, si intreccino tra loro e con i loro partner familiari. Non sono entità fisse, né regole rigide o mezzi di interpretazione localizzabili in funzioni mentali o in pattern neurocognitivi: evolvono nel corso della vita e diventano sempre più precise man mano che l'individuo è esposto alle pratiche narrative.

Dunque, è nella forma di storie che impariamo a strutturare una comprensione più articolata e legata a qualcosa che non possiamo vedere, sentire, percepire in maniera prima di tutto corporea nel comportamento dell'altro. Il nesso con la psicologia del senso comune, cioè con quelle strutture folkpsicologiche che si avvalgono del linguaggio mentalistico per essere espresse, è esplicitato da Hutto proprio a voler sottolineare la non necessità che queste poggino su uno sfondo strettamente legato alla presenza di stati mentali realizzati nell'architettura cognitiva dell'individuo, ma come pezzi di storie di cui facciamo bricolage:

Undeniably, the actions of others sometimes cry out for explanation, but in all such cases, when making sense of these, what we are seeking is a narrative that fills in or fleshes out the relevant details of that person's story. This is the very heart and soul of folk psychological understanding. Hence, I call the narratives that do this kind of work *folk psychological narratives*. The practice of supplying or constructing the *just is* that of explicating and explaining action in terms of reasons. Folk psychology is, by my lights, in essence, a distinctive kind of narrative practice. (Hutto 2008: xi)

Un esempio concreto di *propositional attitude*, che è possibile identificare solamente tramite contenuto proposizionale linguisticamente esprimibile, sono alcuni tipi *credenze*. A differenza di altre caratteristiche intenzionali rintracciabili direttamente nel comportamento dell'agente, come espressioni motorie e facciali, il *mood* affettivo e le manifestazioni di attenzione, credenze come la fede religiosa, posizioni politiche o morali, non sono generalmente immediatamente percepibili nei movimenti dell'individuo nell'ambiente.

Beliefs are a parade example of the kinds of things that lack perceptual signatures. They have no characteristic looks: there is nothing perceptually salient about them; it seems that there is nothing for sensory detectors to get a lock on. [...] We have no direct perceptual experience of the beliefs of others. By this I mean that we do not even "see" the attitude of believing. Beliefs as propositional attitudes are expressed in utterances, not in bodily ways in the way intentional attitudes are. At best, we might perceive these behavioral analogues of beliefs, these nonindividuating attitudes or factual attitudes as expressed in another's outward behavior. (Hutto 2008: 154).

Non riusciamo a comprendere solamente dal riconoscimento delle *intentional attitude* se il nostro interlocutore è, ad esempio, a favore delle teorie cospirazioniste riguardo lo sbarco sulla luna o sulla presenza di una razza aliena che controlla il nostro pensiero, nel senso che non possiamo *riconoscere* la sua azione di aderire a determinate credenze o meno: quello che possiamo riconoscere, invece, sono delle manifestazioni di questo atto di credere, come ad esempio il fatto che il nostro amico cospirazionista indossi un cappellino di carta stagnola. È solamente per queste tipologie di credenze e non altre – come ad esempio la disposizione comportamentale realizzata credendo che il pane si trovi nella mensola della cucina in alto a destra, o che la bambola nel test di Sally e Anne venga cercata nella scatola sbagliata (cfr. Gallagher & Hutto 2008) – che il potenziale di significato non viene afferrato nell'interazione

corporea e situata, introducendo inoltre la possibilità di essere caratterizzate in relazione al proprio contenuto, una volta acquisite le capacità di gestirlo<sup>149</sup> e le abilità di collocarsi all'interno di pratiche socioculturali ben definite<sup>150</sup>.

Questa considerazione sull'atteggiamento proposizionale è supportata anche da una nuova visione del contenuto proposizionale, le cui origini si rintracciano non nella configurazione del processo di recupero e costruzione rappresentazionale dell'informazione dall'ambiente, come negli approcci teleosemantici, ma in ciò che abbiamo introdotto all'inizio di questo capitolo con Haugeland sotto l'etichetta di approccio neo-pragmatista all'intenzionalità:

In line with the neo-pragmatist tradition, RECers maintain that “the primary bearers of content are semantically articulated symbols, occurring in appropriate dynamic patterns” (Haugeland 1990, 412). The job is then to seek to explain how contentful states of mind actually come into being through a process of mastering special kinds of sociocultural practices. (Hutto & Myin 2017: 124).

È, dunque, entrando all'interno dell'ambiente culturale, sociale e collettivo delle istituzioni e delle narrazioni che arriviamo ad acquisire, nel percorso di crescita, la capacità di gestire forme di contenuto proposizionale, la cui origine è da rintracciare nelle forme condivise e vive del linguaggio, e ad impiegarle per costruire storie in cui i protagonisti fanno quello che fanno per delle ragioni.

But how, when, and where does content enter the picture? Content only arises when special sorts of sociocultural norms are in place. The norms in question depend on the development, maintenance, and stabilization of practices involving the use of public symbol systems through which the biologically inherited cognitive capacities can be scaffolded in particular ways. (Hutto 2008: 145)

È fondamentale esplicitare che le *propositional attitude*, una volta consolidate le prime come strumento di interpretazione, non sostituiscono le *intentional attitude* rendendole superflue. Al contrario, le due forme di comprensione si intrecciano a coprire diversi, spesso sovrapposti, aspetti dell'interazione, coniugandosi in processi di costruzione di significato distribuito e multidimensionale: le *intentional attitude* continuano a guidare in modo strutturale il processo di comprensione che avviene in interazione viva con gli altri ma, una volta che la capacità di gestire le *propositional attitude* è acquisita, queste consentono di scalare a un livello di complessità in cui la comprensione include le pratiche socio-culturali, le istanze collettive, le norme e gli usi di una comunità di interpreti, il riconoscimento di contesti formali, e anche a costruire salienze che si riverberano sul livello percettivo (Hipolito et al. 2020). Questo ultimo aspetto risulta fondamentale

---

<sup>149</sup> «Contentful thought only become so when the use of external symbols is mastered». (Hutto 2008: 127)

<sup>150</sup> «[...] it is only when very special forms of sociocultural practices are mastered that the “question of truth” can arise». (Hutto & Myin 2017: 119)

per chiarire il rapporto tra *intentional* e *propositional attitude* e viene ben esemplificato dalla nozione di *affordance sociali*, intesi come la relazione istituita tra aspetti dell'ambiente socio-materiale e le abilità presenti e sfruttabili in una forma di vita, o organismo (Rietveld & Kiverstein 2014; Kiverstein & Rietveld 2021). La costruzione e percezione di *affordance* sociali è un processo cruciale per la comprensione intersoggettiva, la quale si basa i) sull'individuazione del contesto interazionale, ii) sulla capacità di riconoscimento e gestione delle *intentional attitude*, iii) sulla abilità di coniugare *propositional attitude* utili al dispiegamento dell'interazione stessa. In questo senso, l'immersione in pratiche linguistiche che comprendono la lettura, la scrittura, ma soprattutto il discorso – la storia – pronunciato e ascoltato, ovvero l'esercizio individuale e sociale del linguaggio come mezzo di costruzione dei legami interindividuali, costituisce uno dei terreni su cui si incontrano *intentional* e *propositional attitude*.

Among the skills and abilities people develop in the human form of life are skills for expressing, either in the activity of speech or in writing, ways of thinking about the world. Speech and writing are deeply sedimented in the sociomaterial practices in which humans participate, seamlessly interwoven into much of what we do in concrete situations, and often transparently in ways that we as speakers fail to notice. Skills and abilities for expressing things in speech and writing have transformed human life by permeating the ecological niche human inhabit. The affordances of the human ecological niche can thus be said to be “enlanguaged” because they have been formed, at least in part, through the activities of people taking part in sociomaterial practices of speaking and writing. (Kiverstein & Rietveld 2021: 177)

Uno dei terreni, appunto, ma non l'unico. Si pensi, ad esempio, al gesto dell'“occhiolino”. In contesti di interazione sociale, tale espressione può veicolare molti significati – ad esempio intesa o complicità, ironia o sarcasmo, il tentativo di un flirt, il mio compagno di giochi a carte ha l'asso di briscola – ed è solo nell'interazione viva che posso cogliere e riconoscere l'*intentional attitude* del gesto, anche grazie al fatto che ho “assorbito” i significati culturali dell'espressione, largamente usata nella nostra cultura. La comprensione del gesto è immediata, non solo visto che il gesto è chiaro e ben identificabile, ma perché il contesto fornisce le indicazioni attraverso cui sfolire la possibilità di significazione dell'occhiolino e l'interazione stessa mi aiuterà a ritagliare il piano adeguato. L'occhiolino non è quindi solo un movimento, un'azione performata dall'agente intenzionale, ma un'*affordance* sociale che invita all'interazione e alla comunicazione in modi specifici, creando opportunità relazionali che vanno al di là del significato corporeo (strizzare l'occhio) del gesto stesso.

Ora, è chiaro che il cambiamento di prospettiva realizzato tramite lo spostamento verso una concezione enattivista dell'intenzionalità influenzi profondamente il modo di considerare il ruolo da essa svolto nella comprensione interindividuale: l'intenzionalità non viene vista come qualcosa da attribuire tramite inferenze a stati mentali, ma come qualcosa che emerge dall'interazione stessa e che si configura nell'incontro incarnato degli individui.

Questo spostamento, allora, comporta diversi arricchimenti della prospettiva:

- i) Da una parte, introdurre lo strumento delle *intentional attitude* e, in generale, dell'intenzionalità enattivista, apre alla possibilità di considerare l'interazione su un piano diverso, legato alle dinamiche cinestesiche, cinematiche e sensorimotorie, che non viene necessariamente a manifestarsi sul piano linguistico. In questo modo, le modalità attraverso cui primariamente comprendiamo gli altri si manifestano nel nostro modo di rispondere e coordinarci alle *intentional attitude* che rintracciamo nel loro comportamento, nelle espressioni generate in contesti specifici, e che entrano a far parte del computo dell'interazione come fattore non lineare di sviluppo dinamico, nella costruzione della reciprocità come cifra distintiva di un'interazione sociale genuina.
- ii) Inoltre, introdurre gli atteggiamenti intenzionali nel dibattito implica considerare la prospettiva di entrambi gli agenti in interazione: se l'atteggiamento intenzionale alla base degli approcci dell'attribuzione forniva una spiegazione unidirezionale sull'attribuzione dell'intenzionalità all'altro, l'*intentional attitude* enattivista deve essere invece intesa come qualcosa che si manifesta sia nel comportamento dell'agente con cui mi trovo a interagire (anche solamente in forma spettatoriale), sia in me stessa (anche la mia comprensione si manifesta tramite delle *intentional attitude*, ad esempio annuendo all'interlocutore, o nella direzione del mio sguardo), nel tentativo costante di costruire un'armonia insieme all'altro agente<sup>151</sup>. Questo approccio offre così una visione bidirezionale della relazione interpersonale, in cui la comprensione non è un percorso a senso unico, ma si nutre delle rispettive forme di intendimento e crea una dimensione autonoma in cui è possibile co-regolare i significati messi in gioco.
- iii) Dall'altra parte, la considerazione delle *propositional attitude* ci consente di aprire a uno spazio di interpretazione che è nutrito dalle pratiche linguistiche, le quali a loro volta fanno affidamento a un sostrato incarnato, percettivo e sensorimotorio, ma che sviluppano significati nel dominio dell'astrazione, dell'assenza, dell'immaginazione, della virtualità narrativa dei possibili effetti pratici. In questo senso, è essenziale riconoscere come le *propositional attitude* non siano semplicemente un'addizione al discorso enattivista, ma rappresentino un ponte critico verso la comprensione delle interazioni più complesse che caratterizzano le dinamiche sociali. Esse ci permettono di esplorare come le intenzioni e i

---

<sup>151</sup> È importante precisare che la costruzione dell'armonia interindividuale può avvenire sia attraverso aspetti di rinforzo positivo dell'interazione, cercando di assecondare i movimenti e le intenzioni dell'interlocutore per condurre l'interazione verso una forma congiunta di intenzionalità, sia nel caso in cui si desideri in qualche modo ostacolarla. In altre parole, la descrizione che abbiamo fornito non vale solo per contesti collaborativi, pacifici e sincronici: anche per *disarmonizzare* l'interazione, infatti, è fondamentale prima di tutto comprendere l'altro.

significati si articolano e si negoziano all'interno di contesti comunicativi specifici, permettendo agli agenti di operare con una flessibilità che va oltre la reattività fisica.

- iv) Inoltre, la considerazione delle *propositional attitude* offre una prospettiva arricchita sui meccanismi di co-costruzione del significato, evidenziando come la comunicazione non sia mai identificabile con un semplice scambio di informazioni, ma si caratterizzi meglio se intesa nei termini di un atto creativo che emerge dall'intreccio di esperienze vissute, intuizioni e prospettive dell'individuo coinvolto. In tal modo, si instaura un dialogo ricco e dinamico in cui ogni agente diventa co-autore del significato, contribuendo attivamente a una comprensione condivisa che si evolve e si adatta nel tempo. Questa dimensione relazionale dell'interazione linguistica non può essere sottovalutata, poiché riflette la complessità della condizione umana e la nostra capacità di costruire mondi condivisi secondo almeno un doppio canale: il linguaggio e l'azione.
- v) L'atteggiamento intenzionale, dunque, si intreccia con quello proposizionale in situazioni che richiedono strumenti complessi per coordinarsi e comportarsi in maniera adeguata. Le *propositional attitude* si costituiscono come strumenti linguistici e narrativi a cui posso ricorrere quando la modalità di comprensione più immediata non riesce a esaurire il senso del comportamento dell'altro: ad esempio, quando le mie aspettative sono fortemente disattese, quando devo ragionare con un amico sul perché la sua ragazza lo ha lasciato, o quando devo elaborare una strategia logica per prevedere il comportamento dell'altro. Tuttavia, come abbiamo visto, il confine tra atteggiamento intenzionale e proposizionale non è rigido: una volta acquisita la capacità di gestire gli atteggiamenti proposizionali, questi due modi di intendere la cognizione possono mescolarsi in maniera imprevedibile, influenzati dall'esperienza individuale, dal contesto culturale e dalle inclinazioni personali.

Per i motivi che abbiamo appena elencato, che delineano una visione completamente diversa del ruolo che l'intenzionalità assume nei contesti intersoggettivi rispetto agli approcci precedentemente analizzati, ci sembra appropriato suggerire un cambiamento nel modo di definire il fenomeno che vede l'intenzionalità protagonista nella comprensione interazionale. Proponiamo, quindi, di sostituire la dicitura "attribuzione di intenzionalità", che classicamente denota il processo di comprensione dell'altro come agente intenzionale tramite ascrizione di stati mentali, con la terminologia *intentionality detection*, o "riconoscimento/individuazione/rilevamento dell'intenzionalità". Mentre la prima visione implica una concezione spettatoriale dell'incontro con l'altro, in cui il ruolo del corpo e dell'interazione viene del tutto trascurato, parlare di *detection* dell'intenzionalità, invece, ci consente di riconoscere che la comprensione dell'intenzionalità altrui si fonda sull'afferrare la

struttura del comportamento esibito dall'intero organismo, senza necessità di astrazioni cognitive o rappresentazioni mentali. Allo stesso tempo, se l'attribuzione comporta una fortissima componente razionale nel processo di decisione di quale *stance* adottare o quale stato mentale attribuire, la *detection* si configura come una modalità primariamente corporea di comprensione, che eventualmente può riconfigurarsi in *propositional attitude* ma in cui il legame tra le due non è garantito a priori da un isomorfismo ontologico o da una scommessa sulla traducibilità della struttura della cognizione nel linguaggio impiegato per descriverne i risultati. Nella prospettiva inaugurata dall'*intentionality detection*, quindi, viene garantita la possibilità di descrivere il processo di comprensione dell'altro in termini intenzionali sotto due diverse prospettive: quella corporea, immediata, visibile nei comportamenti di *attunement* interazionali a loro volta intenzionali e quella narrativa, linguistica, che può prevedere anche l'uso di termini mentalistici. Così, i processi attitudinali possono essere visti come distinti e anche scoordinati, dal momento che le *propositional attitude* possono non essere in relazione con le *intentional attitude*.

Siamo partiti da un approccio che teneva insieme due idee eterogenee sull'intenzionalità, di cui abbiamo cercato di mostrare le incompatibilità e le conseguenze nei contesti sperimentali che sopravvengono al momento dell'interpretazione dei dati. Cosa succede, quindi, se proviamo ad interpretare i risultati degli esperimenti realizzati da Marchesi e colleghi (2019, 2021a) adottando la proposta enattivista dell'*intentionality detection*?

#### 4.3.3 Un'analisi dell'*intentionality detection* nella HRI

Alla fine del paragrafo §4.2.3, avevamo posto degli obiettivi, fissati al fine di rendere conto delle incongruenze teoriche e interpretative riscontrabili negli esperimenti sull'attribuzione di intenzionalità ai robot sociali. In particolare, i) ci siamo proposti di fornire una spiegazione dell'*intentional stance* per come formulata da Dennett (1971, 1987), ovvero che non sia naturalizzata in termini di processi cerebrali o che faccia appello a derivati del *Language of Thought* interno alla mente, preservandone invece la natura esterna come strategia linguistica e narrativa alimentata dalla psicologia del senso comune. Il secondo obiettivo ii) consisteva nella necessità di gestire una situazione in cui, negli esperimenti per valutare l'attribuzione di intenzionalità ai robot, emergono risultati spesso contraddittori a livello personale e subpersonale, rispettivamente misurate esplicitamente e implicitamente. Vediamo, ora, se abbiamo esaudito i nostri propositi.

Per quanto riguarda i), l'*intentional stance* sembra perfettamente compatibile con la definizione di *propositional attitude*. Infatti, possiamo connotare l'approccio delle *attitude* sviluppato da Hutto come un'ibridazione, con alcune variazioni epistemologiche, dell'approccio neo-comportamentista fornito da Dennett e di quello neo-pragmatista che individua l'origine del significato nelle pratiche

condivise e normate dalla collettività. Nel dettaglio, le *propositional attitude* condividono con l'*intentional stance* la funzione di rendere intelligibili le azioni e le credenze di un agente in base a una comprensione complessa del comportamento; con l'approccio neo-pragmatista, condividono la natura sociale, pragmatica e narrativa dell'origine del contenuto intenzionale. Le variazioni epistemologiche riguardano il sostrato teorico: mentre le *propositional attitude* si concentrano su contenuti esplicitamente articolati in proposizioni — come credenze o desideri che sono formulabili linguisticamente — l'*intentional stance* di Dennett non implica necessariamente una formalizzazione proposizionale. In questo senso, l'*intentional stance* si basa su un estremo razionalismo e logicismo dell'agente che lo adotta, anche se la verbalizzazione non avviene in termini di contenuti proposizionali. Hutto, d'altra parte, pur accettando una parte di questa dinamica interpretativa, si distanzia dall'idea che tutte le forme di atteggiamento intenzionale debbano essere rappresentate attraverso proposizioni, e scorpora l'impiego delle *propositional attitude* da forme più corporee e immediate di comprensione intersoggettiva. Nella sua visione, le intenzioni e gli scopi possono emergere da contesti pratici condivisi senza necessariamente richiedere una struttura proposizionale. Pertanto, l'*intentional stance* di Dennett può essere considerata compatibile su un piano teorico alle *propositional attitudes*, a patto di non renderla l'unica modalità di comprensione dell'altro e porre, invece, maggiore enfasi sulle dinamiche relazionali e pratiche che originano dalle interazioni sociali vive. In questo caso, l'ibridazione non comporta contraddizioni sostanziali, dato che entrambi gli approcci condividono una visione externalista circa le origini dell'intenzionalità, così come proposto nella definizione delle *propositional attitude* tramite l'esplicito richiamo all'approccio neo-pragmatista per come spiegato da Haugeland (1990). Quindi, l'atteggiamento intenzionale di Dennett può agevolmente essere incluso nella serie di *propositional attitude* che nutrono le pratiche narrative attraverso cui diamo senso alle azioni degli altri in contesti di interpretazione più complessi. A rinforzare la compatibilità, inoltre, va annoverato il richiamo comune alla *folk psychology* come bacino di spiegazioni e motivazioni in cui l'individuo forma la propria capacità predittiva e interpretativa e a cui attinge per elaborare gli atteggiamenti proposizionali. Sarebbe un errore, dunque, pensare all'*intentional stance* come una forma di *intentional attitude*: l'atteggiamento intenzionale di Dennett, in quanto forma astratta delle strategie adottate per fornire spiegazione del comportamento degli altri che fa uso del linguaggio della psicologia del senso comune per costruire descrizioni basate sul linguaggio mentalistico, è invece una *propositional attitude*.

Se applichiamo tale riflessione agli esperimenti di Marchesi et al. (2019, 2021a), riusciamo a vedere come la situazione possa essere interpretata in modi molto diversi da quanto fatto dai ricercatori. Quando i collaboratori chiedono ai soggetti dell'esperimento (Marchesi et al. 2019) di scegliere tra descrizioni che contengono termini meccanici o mentali, non stanno trovando evidenze



per l'*intentional attitude*, ma stanno testando le *propositional attitude*. Non è il tipo di intenzionalità corporea, immediata, situata che stanno interrogando, ma una modalità più astratta di dare senso agli altri, la quale non è del tutto disimplicata da una modalità più profondamente incarnata, ma è comunque differente. Questa idea sembra chiara in Marchesi et al. (2019), in cui i ricercatori hanno correttamente progettato un questionario per valutare se i partecipanti adottassero l'atteggiamento intenzionale di fronte ai robot, poiché se ciò che stanno cercando è capire come i partecipanti spiegano linguisticamente e razionalmente il comportamento degli altri, allora un questionario che mette a confronto i diversi atteggiamenti tramite l'uso di linguaggi diversi è lo strumento adeguato a scoprirlo. Tuttavia, questi risultati non dimostrano affatto che i partecipanti non comprendano del tutto i robot come agenti intenzionali, ma solamente che le *propositional attitude* che vengono utilizzate fanno principalmente affidamento a termini che identificano il robot come un agente interpretabile tramite strategia progettuale, ovvero non impiegando termini mentalistici. Anzi, come ha mostrato anche l'esperimento successivo (Marchesi et al. 2021a), la replicazione dei risultati registrati nello studio del 2019 conferma la possibilità – anche se meno frequente in un campione significativo – di adottare l'*intentional stance* nei confronti del robot, ma solo se intesa nella sua forma proposizionale di impiego di contenuti mentalistici collezionati nel bagaglio personale di pratiche narrative con cui diamo senso al mondo in modo linguistico. Come interpretare, invece, i risultati impliciti raccolti tramite la misurazione dei tempi di risposta?

Arriviamo quindi a ii), ovvero ad affrontare una situazione in cui le misurazioni implicite producono evidenze di una plausibile tendenza a adottare atteggiamenti intenzionali, mentre a livello esplicito si registra una netta preferenza verso le attitudini meccaniche. Ricordiamo che, nel caso che stiamo analizzando, i ricercatori avevano individuato la presenza di tempi di risposta commisurabili e non troppo divergenti nella scelta di descrizioni mentalistiche e meccaniche per lo scenario con il robot, che li ha fatti propendere verso un'interpretazione che individua sia l'*intentional* che la *design stance* come «compatible with participants' mindset, at least at the implicit level» (Marchesi et al. 2021a). Alla luce dei cambiamenti che abbiamo proposto, ora siamo in grado di riconoscere che i due tipi di misurazione impiegati in questo studio inquadrano entrambi comportamenti che non sono legati alla comprensione corporea, ma alla considerazione del contenuto proposizionale veicolato dalle *propositional attitude*. Infatti, il metodo di misura esplicito, tramite il questionario, inquadra direttamente la presenza dell'*intentional* o della *design stance*, mentre la misurazione dei tempi di risposta correla la comprensione del contenuto proposizionale della descrizione in calce alla sequenza fotografica a un meccanismo di accettazione o di rifiuto della *propositional attitude* proposta. Se adottiamo l'approccio enattivista che abbiamo prospettato, non si può considerare questo come un tentativo di vagliare forme implicite di intenzionalità del tipo identificabile tramite le *intentional*

*attitude*. Anche se, come accade, si registrano velocità differenti ma commensurabili per l'accettazione delle descrizioni che includono termini mentalistici e termini meccanici, secondo la visione che proponiamo tale evidenza non può essere considerata afferire a forme corporee di intenzionalità, dal momento che il compito assegnato richiede all'individuo una comprensione linguistica che passa per la lettura e la validazione delle descrizioni offerte e che concerne l'accordo o il disaccordo del contenuto proposizionale veicolato dalla *propositional attitude*. Vedendola sotto questa prospettiva, la similarità registrata nei tempi di risposta tra descrizioni mentalistiche e meccaniche nello scenario con iCub può essere considerata un ulteriore rinforzo dei risultati provenienti dall'IST, ma non fornisce prove più implicite circa la considerazione del robot come agente intenzionale.

Ora, se la registrazione dei tempi di risposta non costituisce evidenza del riconoscimento del robot come agente intenzionale in termini enattivisti, come fare a verificare se l'essere umano protagonista dell'esperimento rileva un comportamento implicitamente intenzionale quando si trova di fronte al robot?

Innanzitutto, considerando in dettaglio i principi enattivisti presentati in questo capitolo e nel precedente, il contesto sperimentale deve tenere conto di almeno due fattori:

- a) In primo luogo, è essenziale che lo scenario implichi la presenza fisica di entrambi gli agenti, quello umano e quello robotico, in modo da poter istaurare dinamiche interazionali e avere la possibilità di raccogliere, rilevare e interpretare in modo incarnato i movimenti del robot. Infatti, negli esperimenti di Marchesi et al. (2019, 2021a) che abbiamo analizzato, il contesto sperimentale disincarnato, allestito tramite uno schermo e che prevede la decisione senza possibilità di intervenire in alcun modo su quello che viene presentato, potrebbe influenzare in modo sostanziale la comprensione del robot come agente intenzionale. Questa possibilità è resa ancora più concreta dal fatto che l'agente robotico protagonista dell'esperimento, iCub, non è nemmeno in movimento. Anzi, è l'individuo che, alla presentazione della serie di immagini, deve operare un'integrazione affinché l'azione si "svolga" davanti ai suoi occhi, aggiungendo un ulteriore livello di astrazione. Abbiamo visto che, nell'impianto teorico a sostegno dell'esperimento, questo è giustificato dall'adesione ai principi della ToM, in cui l'interazione viva non è considerata indispensabile alla comprensione dell'altro e anzi si concentra sulle dinamiche interne alla mente che avvengono, in sostanza, in solitudine sia cognitiva che corporea. Possiamo, ora, serenamente sorpassare questa visione in favore della prospettiva enattivista alla cognizione sociale: se di interazione tra umani e robot stiamo parlando, allora la sua indagine sperimentale non può basarsi sulla semplice osservazione di immagini del robot, ma deve includere il livello corporeo e incarnato, che permette di *sentire*

la presenza del robot e di costruire cicli sensorimotori di aggiustamenti pragmatici reciproci. Questa sensibilità è, d'altronde, sempre più al centro del dibattito sulla HRI, e si delinea come una delle questioni più dirimenti al momento nel campo di studi (Woods et al. 2006; Sciutti & Sandini 2017; Sheridan 2016; Lindblom et al. 2020; cfr. §4.4.2).

- b) Seguendo l'approccio che abbiamo proposto, testare se l'individuo comprende il robot in interazione come un agente intenzionale significa esplorare i modi in cui, a livello corporeo e di comportamento, l'essere umano si coordina, reagisce, si connette alle azioni performate dal robot e al comportamento che questo esibisce in interazione. Si tratta di osservare se, nel duo interazionale che si configura, i comportamenti performati dal robot innescano dinamiche corporee di risposta e aggiustamento nell'essere umano che sono riscontrate in altre occasioni in cui è evidente l'instaurazione di un regime interazionale collaborativo e intersoggettivo. Il corrispettivo da prendere in considerazione è, ovviamente, l'interazione intersoggettiva tra umani: dal momento che vogliamo comprendere come realizzare forme di collaborazione e interazione più efficienti portate avanti da agenti artificiali, prendere ad esempio l'interazione che, nella storia dell'umanità, è stata fino ad ora quella che ha portato risultati migliori di coordinazione e reciprocità di comprensione potrebbe dotarci di un robusto termine di paragone, anche se notevolmente ambizioso. Ciò non significa, d'altra parte, postulare che la HRI debba essere esattamente coincidente, per forma, struttura e reazioni registrate, a quella tra umani, ma mettere in relazione le due consente di vedere differenze magari rimaste ai margini, *invisibili*, soprattutto se si imposta l'indagine come volta a indagare gli aspetti corporei e situati. L'obiettivo si fa interessante: se riscontrassimo delle evidenze della presenza di una considerazione intenzionale sul piano corporeo, queste potrebbero funzionare da punto di partenza per costruire un approccio enattivista alla HRI che descriva le dinamiche interazionali non solo facendo affidamento alla componente esplicita del giudizio sull'interazione stessa da parte dell'umano, ma anche analizzando le maniere in cui il senso condiviso si aggrega e si disgrega, si costruisce per poi sfaldarsi ed essere recuperato o andare totalmente a dissolversi. Troveremo tracce di questo tortuoso percorso del senso reciproco nel modo in cui le azioni e le reazioni si concatenano nell'incontro con il robot.

Si può annoverare una già nutrita serie di esperimenti che evidenzia, in diversi modi, il fatto che l'essere umano interagisca fisicamente con i robot in modi che, anche se minimi, sono comparabili al modo in cui risponde quando è impegnato con un altro essere umano. In questi studi, movimenti particolari, anche minimi, realizzati dall'essere umano di fronte al robot vengono rilevati a partire da una serie di testimonianze consolidate negli studi di cognizione sociale circa il coinvolgimento

corporeo in contesti interazionali. In particolare, le misure che crediamo siano più significative riguardano il ruolo dello sguardo, da una parte, e l'attivazione di alcuni pattern neurocognitivi legati alla componente motoria e alla comprensione intersoggettiva, dall'altra.

Il primo gruppo di studi ha investigato il modo in cui lo sguardo assume un ruolo fondamentale nella comprensione dell'altro in interazione sociale. Infatti, a partire da alcune ricerche che individuano lo sguardo come uno dei canali di comunicazione privilegiati nell'incontro con altri esseri umani, alcuni nuovi metodi sperimentali sono stati importati nella HRI per considerarne l'importanza anche nell'interazione con i robot. In particolare, diversi studi sono stati condotti sulla importanza e sul valore dello sguardo reciproco (*mutual gaze*, Kompatsiari et al. 2017, 2021; Marchesi et al. 2023), sulla direzionalità dello sguardo come indicatore dell'attenzione congiunta (*shared attention*, Pereira et al. 2020; Moon et al. 2014) e sulla sua componente deittica o referenziale (*referential* o *deictic gaze*, Yilmaz et al. 2020; Saran et al. 2018a; 2018b)<sup>152</sup>. L'analisi di uno studio in particolare può farci rendere conto di quanto questo tipo di indagine possa rivelare sulla possibilità che il robot venga inteso come agente intenzionale in interazione con l'essere umano.

Nello studio di Sciutti et al. (2013), i ricercatori hanno impiegato un metodo sperimentale molto diffuso negli studi neurocognitivi sulla cognizione sociale, ovvero l'*anticipatory gaze*. Anche se lo studio parte da presupposti teorici coincidenti con quelli della ToM e dell'attribuzione di intenzionalità, il modo in cui è allestita la situazione sperimentale ci consente di utilizzare i risultati a supporto della nostra tesi sulla comprensione primariamente corporea del robot come agente intenzionale in interazione. Infatti, tecniche di misurazione del comportamento dello sguardo sono largamente impiegate in letteratura in quegli studi che cercano di indagare i meccanismi di comprensione intersoggettiva soprattutto nei bambini in età prelinguistica (Falck-Ytter et al. 2006; Kanakogi & Itakura 2011; Rosander & von Hofsten 2011). In particolare, la natura anticipatoria dello sguardo risulta, tra tutte le altre modalità, quella più interessante e pertinente, dal momento che è stato dimostrato che «when humans observe others performing goal directed actions, they shift their gaze to the target of their movement before it is completed» (Sciutti et al. 2013: 330). Questo tratto anticipatorio dello sguardo nell'essere umano sembra essere indicativo delle capacità di comprendere o immaginare l'obiettivo dell'azione svolta dall'altro, implicitamente verificando, quindi, la considerazione dell'agente come intenzionale. Capire se lo stesso sguardo anticipatorio si realizza nell'essere umano quando si trova in presenza davanti al robot che compie l'azione costituirebbe una prova in tal senso. Con questo scopo, Sciutti e colleghi (2013) adattano il *setting* sperimentale già validato in Falck-Ytter et al. (2006): il partecipante dell'esperimento osserva un attore mentre svolge

---

<sup>152</sup> Per una *review* sugli studi che indagano il ruolo dello sguardo in HRI, si veda Admoni & Scassellati (2017). Per una panoramica completa sul legame tra *mutual gaze* e *joint action*: Marchesi et al. (2023).

il semplice compito di spostare un oggetto da una posizione all'altra su un tavolo di lavoro, sostituendo però l'attore umano con il robot iCub (Metta et al. 2010). Un sistema di *eye tracking* registra i movimenti oculari dell'osservatore al fine di tracciare il comportamento dello sguardo e verificare la presenza di un movimento anticipatorio dell'obiettivo dell'azione. I movimenti del robot sono stati calibrati su una velocità e una traiettoria ricavati dalla mappatura dello stesso movimento performato da un essere umano, così da rendere l'azione più fluida e antropomorfa possibile (cfr. *infra* §4.4.2). Le ipotesi formulate sono due, mutualmente esclusive: il soggetto umano potrebbe implicitamente riconoscere il robot come agente, mostrando quindi l'abilità di anticipare con lo sguardo l'obiettivo dell'azione performato dal robot, oppure, alternativamente, il robot potrebbe essere percepito come un oggetto molto complesso che si muove, ma non come un agente le cui azioni sono orientate a uno scopo e, potremmo aggiungere noi, pervase di *intentional attitude*. In questo secondo caso, l'attività di anticipare l'obiettivo tramite lo sguardo calerebbe drasticamente, mostrando invece «a stronger tendency to track the moving object» (Sciutti et al. 2013: 333). Lo scenario prevedeva, inoltre, una condizione di controllo in cui l'azione veniva realizzata da uno sperimentatore umano, ma sempre dopo che la stessa azione era stata svolta da iCub, così da evitare effetti che avrebbero potuto condizionare il soggetto a interpretare il movimento del robot allo stesso modo di come fatto nella condizione con l'essere umano. In più, onde evitare la possibilità di sfruttare diverse *affordance* sociali, i volti dello sperimentatore e di iCub sono stati nascosti tramite un divisorio.

Arrivando ai risultati, lo studio riscontra che le stesse attività di sguardo anticipatorio sono riscontrabili sia nel primo scenario con iCub, sia nel secondo scenario con lo sperimentatore umano. In altre parole, nella maggior parte dei casi<sup>153</sup>, l'individuo anticipa con lo sguardo l'obiettivo dell'azione di trasporto dell'oggetto sia se svolta dal robot, sia da un altro essere umano: «the [robot] is “motorically” interpreted as an agent» (Sciutti et al. 2013: 341). In questo studio, troviamo allora l'opportunità di mettere al vaglio l'ipotesi delle *intentional attitude*, dal momento che è proprio la risposta corporea a un comportamento esibito che viene testata e che risulta compatibile con la nostra proposta sia nelle modalità di indagine sperimentale, sia nel linguaggio utilizzato per interpretare i risultati. Analizzando il modo in cui l'essere umano interpreta a livello corporeo e motorio il robot, ciò che possiamo trovare è una risposta diretta, immediata e conforme a quelle che si verificano in altre interazioni sociali che l'individuo si trova a costruire quando incontra altri agenti intenzionali. Il vantaggio di studi come questo consiste in almeno tre aspetti:

---

<sup>153</sup> Nei casi in cui l'individuo non ha mostrato comportamenti di *anticipatory gaze* nei confronti del robot, gli autori hanno controllato lo scenario in cui era l'umano a performare l'azione e hanno riscontrato che l'individuo non mostrava il comportamento visivo anticipatorio nemmeno in questo caso. Il risultato è interpretato così: «This suggests that the disappearance of anticipation was not due to the presence of a robotic artefact as demonstrator but rather to other factors, which modulate gazing behaviour also during human action observation». (Sciutti et al. 2013: 339)

- i) Non cerca di naturalizzare meccanismi espliciti di attribuzione di intenzionalità, rintracciando correlati interni di comportamenti manifestati in corrispondenza di giudizi linguistici, non ipotizzando quindi che la comprensione corporea si appoggi a una struttura linguistica di base e, anzi, situando i livelli più basilari e fondamentali della comprensione intersoggettiva proprio a livello corporeo e incarnato.
- ii) Introduce all'interno del dibattito sul ruolo dell'intenzionalità una tecnica di misurazione delle risposte implicite che è ben documentata negli studi di cognizione sociale, la quale si basa su un modello teorico *embodied* della comprensione interpersonale e che potrebbe fornire la base per uno sviluppo più approfondito della validazione delle risposte comportamentali motorie nel processo di comprensione dei robot in interazione. Tale introduzione è in linea con l'appello espresso in più sedi circa la necessità di importare all'interno degli studi di HRI delle misure sperimentali provenienti da altri domini, che possano arricchire l'indagine sull'interazione tra umani e robot facendo affidamento a tecniche consolidate in studi paralleli (cfr. Perez-Osorio & Wykowska 2019).
- iii) Lo studio viene condotto in presenza, con il robot e l'individuo umano che condividono uno spazio di possibilità interazionali. È chiaro che, a causa della fragilità dell'artefatto robotico e della necessità di garantire la validità dei risultati e la replicabilità dello studio, contesti in cui avviene una effettiva interazione, magari multisensoriale, sfruttando il tatto, ad esempio, o improvvisata sono molto difficili da realizzare, anche se sicuramente costituiscono uno degli obiettivi a medio o lungo termine. Ciò che è fondamentale, tanto per il nostro approccio quanto per gli autori dello studio, è che l'incontro tra le parti avvenga in un contesto interazionale vivo. Questo aspetto è reso pertinente anche dagli stessi autori, che tengono a sottolineare l'importanza di allestire l'esperimento in presenza:

It should be also noted that the robot was physically present, executing the action in front of the subject's eyes. Such concrete presence made it clear for the subject that the robot was a real artefact and just an animated character or a human actor in disguise, which could have in turn possibly evoked a response similar to human observation by analogy. (Sciutti et al. 2013: 343)

Possiamo prendere in esame un ulteriore aspetto di questo esperimento che riteniamo concorde alla visione enattivistica proposta per introdurre l'altra misura implicita che può essere impiegata per indagare la presenza di *intentional attitude* nel comportamento umano di fronte al robot. Infatti, lo studio si struttura su una serie di indagini riguardo alla *motor resonance* nella cognizione sociale, ovvero l'attivazione automatica del sistema motorio di chi percepisce un'azione mentre questa si svolge (Rizzolatti et al. 1999) come misura neurofisiologica affidabile per valutare il coinvolgimento

di un individuo in un contesto di interazione sociale. Nella discussione in cui vengono interpretati i risultati dello studio, gli autori sottolineano in modo esplicito il legame tra sguardo anticipatorio e una componente neuromotoria che garantisce la corrispondenza corporea e l'allineamento nell'azione degli agenti coinvolti:

In summary, during the observation of a humanoid robot performing a goal directed action as transporting an object into a container, subjects anticipatorily gaze to the goal of its action the same way they would for the presentation of a human action, suggesting that the robot has been implicitly interpreted as a goal-oriented agent and not just as a complex moving object. This occurrence of anticipatory gaze shift implies a motor-based understanding of the action goal, which does not require inferential or teleological reasoning, but is rather based on an implicit, covert matching between the observer's and the agent's actions. (Sciutti et al. 2013: 344)

Il legame che gli autori stringono tra rilevazione del comportamento anticipatorio dello sguardo e risonanza motoria in questo studio proviene da un altro articolo, di un anno precedente. In quella sede, Sciutti e colleghi (2012a) si pongono l'obiettivo di «offer a different perspective in suggesting that human responses can become a tool to measure and improve robot interactional attitudes» (2012a: 223) e fanno riferimento agli studi sulla risonanza motoria come strumento per capire se il robot esibisce atteggiamenti interazionali. Affermano che altre forme di verifica implicite utilizzate nei contesti sperimentali<sup>154</sup>, infatti, come la misurazione della frequenza cardiaca o la *galvanic skin conductance*, non sembrano essere strettamente correlate ai processi al cuore della comprensione interindividuale, mentre la risonanza motoria coglie proprio uno dei meccanismi centrali alla base della cognizione sociale, ovvero l'accoppiamento tra azione e percezione dell'altro in interazione (Sciutti et al. 2012a: 224). Le modalità attraverso cui viene descritta la nozione di *motor resonance* sono indicative della vicinanza con l'approccio enattivista che abbiamo proposto:

The word “resonance” has been chosen because, as two identical and close diapasos vibrate together when one of the two starts its vibration, during action observation the two motor brains “resonate” because they share a similar motor repertoire. (Sciutti et al. 2012a: 224)

Nell'articolo, inoltre, gli autori annoverano due principali modalità di manifestazione della risonanza motoria: un comportamento proattivo dello sguardo, o *proactive gaze behavior*, di cui l'*anticipatory gaze* messo alla prova nello studio del 2013 è una declinazione, e il comportamento di imitazione automatica, o *automatic imitation behavior*. È questa seconda modalità di testare la risonanza motoria che il gruppo di ricercatori applica nello studio di Bisio et al. (2014), in cui cercano di «investigate how the nature of the observed agent, the object-directedness, and the kinematic properties of the observed movement influence a human observer when interacting with an external

---

<sup>154</sup> Per una rassegna delle misure neurofisiologiche e neurocognitive in HRI, si veda rispettivamente Wiese et al. (2017) e Henschel et al. (2020); Cross et al. (2019).

agent» (Bisio et al. 2014: 2). In particolare, lo scopo è individuare se è possibile innescare meccanismi di *motor contagion* nell'essere umano di fronte al robot nello stesso modo in cui succede in interazione con altri esseri umani. Infatti, la risonanza motoria che si verifica come processo implicito e non inferenziale di comprensione dell'altro ha come possibile effetto la creazione di processi imitativi non coscienti, in cui l'osservatore ripete l'azione svolta dall'altro copiando il movimento che ha appena visto compiersi. Nelle parole di Bisio e colleghi, «if the motor system is prepared to produce a motor response to the observed motion, this might result in motor contagion: namely, the observer's motor performance might automatically replicate some features of the stimulus» (2014: 1).

A partire da questo connubio, i presupposti teorici sviluppati vengono messi alla prova in un esperimento che prevede che i partecipanti osservino separatamente un robot umanoide e un essere umano eseguire un'azione che può essere di tipo transitivo (diretta verso un oggetto, ovvero riporre un oggetto in un contenitore) o intransitivo (senza oggetto, cioè muovere le mani), la quale può essere svolta attraverso movimenti fluidi e accompagnati o bruschi e artificiali. Nella condizione con il robot è inclusa anche la possibilità che questo si muova in maniera non conforme a velocità biologiche, ovvero con accelerazioni e interruzioni improvvise, ad esempio. Il compito prevede poi che i partecipanti ripetano il movimento osservato mentre gli sperimentatori misurano quanto la velocità del movimento dei partecipanti viene influenzata dalla velocità con cui l'azione osservata è stata compiuta dal robot o dall'umano, verificando quindi l'eventuale presenza di un contagio motorio. Le ipotesi formulate riguardano la conformità dell'imitazione della velocità dell'azione da parte del soggetto osservatore (H1) e la possibile influenza esercitata dalla presenza di un oggetto nell'azione (H2):

- (H1): se il meccanismo di risonanza motoria dipende principalmente dalla corrispondenza tra il repertorio motorio dell'osservatore e quello del dimostratore (sia un essere umano o un robot), allora il contagio motorio si manifesterà indipendentemente dalla natura dell'agente osservato (che sia umano o robot). In altre parole, se l'osservatore può riconoscere nel movimento del robot o dell'umano un repertorio motorio simile al proprio, si verificherà un'imitazione della velocità del movimento dell'agente, indipendentemente dalla forma della traiettoria (fluida-curvilinea o brusca-artificiale). Tuttavia, quando il robot esibisce un movimento con cinetiche non biologiche, cioè movimenti che non corrispondono a quelli che potremmo aspettarci da un corpo umano (per esempio, movimenti innaturali o robotici), allora la risonanza motoria non dovrebbe essere attivata. Nel caso in cui il contagio motorio non si verifichi, la velocità dei partecipanti non sarà influenzata dal movimento osservato.



- (H2): se la presenza di un oggetto come scopo dell'azione induce i partecipanti a concentrarsi principalmente sul raggiungimento dell'obiettivo (azioni transitive) e non sul modo in cui l'azione viene eseguita (cinetica del movimento), allora la velocità con cui i partecipanti eseguono l'azione non sarà influenzata dalla velocità del dimostratore.

I risultati registrano piena conformità alle ipotesi formulate, «show[ing] that the observed movement kinematics, namely an action property at low cognitive value, modulated the resonance mechanisms, whatever the nature of the observed stimulus, except when the humanoid robot violated the biological law of motion» (Bisio et al. 2014: 7). Nello specifico, H1 è verificata poiché i risultati dimostrano che il contagio motorio si è manifestato sia quando i partecipanti osservavano un agente umano, sia quando osservavano il robot umanoide, a condizione che i movimenti del robot rispettassero le cinetiche biologiche. Questo conferma che la corrispondenza tra il repertorio motorio dell'osservatore e quello dell'agente è un fattore cruciale per l'attivazione dei meccanismi di risonanza motoria: quando i partecipanti erano in grado di riconoscere i movimenti del robot come simili a quelli che si aspettano da un agente biologico, la velocità in cui veniva imitata l'azione si sincronizzava a quella del robot, confermando così l'ipotesi che il contagio motorio si verifica indipendentemente dalla natura dell'agente osservato (umano o robot), purché il movimento segua leggi biologiche. Tuttavia, come previsto, quando il robot eseguiva movimenti non mantenendo velocità costante o esibendo cambiamenti innaturali di accelerazione, il contagio motorio non si è manifestato, a dimostrazione che la risonanza motoria non viene attivata se il movimento osservato è percepito come innaturale o estraneo al repertorio motorio dell'essere umano. Anche H2 viene confermata, poiché i risultati mostrano che la presenza di un oggetto come scopo dell'azione modifica il comportamento dei partecipanti. In particolare, quando l'azione era transitiva (ossia diretta a un oggetto), l'influenza della cinetica del movimento osservato sui partecipanti è risultata ridotta. Questo suggerisce che, quando l'obiettivo dell'azione era chiaramente definito, i partecipanti tendevano a concentrarsi principalmente sul raggiungimento dell'obiettivo e non tanto sul modo in cui l'azione veniva eseguita. In altre parole, la cinetica del movimento aveva un impatto minore nelle condizioni transitive, supportando l'H2.

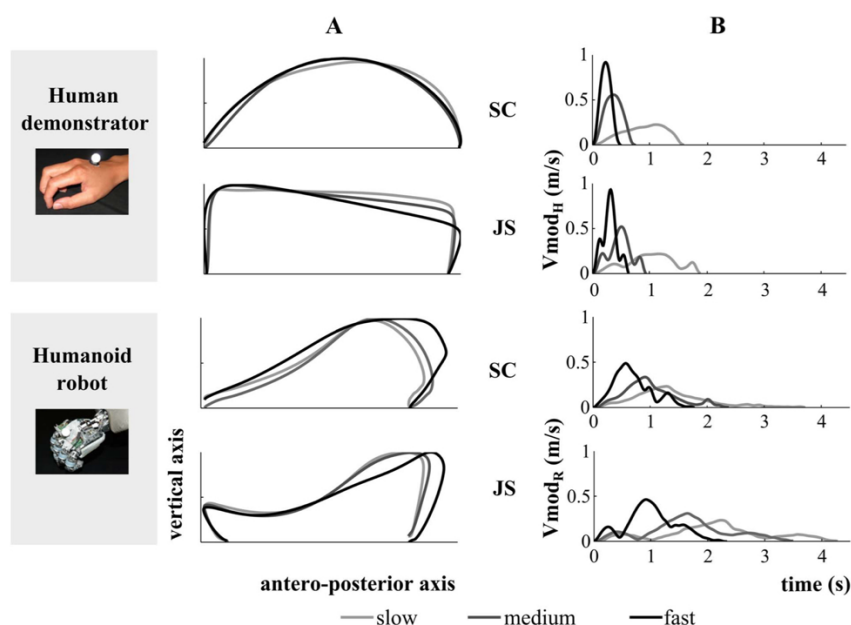


Fig. 4.3 Esempi di traiettorie e profili di velocità quando i dimostratori realizzano movimenti biologici. Sulla sinistra, sono rappresentate le traiettorie dei dimostratori umano e robotico sul piano sagittale; sulla destra, i moduli dei profili di velocità dei due stimoli. Adattato da Bisio et al. (2014: 4).

Dunque, anche il contagio motorio, in quanto forma di *motor resonance*, può essere annoverato tra le modalità di indagine delle *intentional attitude* dell'essere umano di fronte al robot. A rinforzare ulteriormente la compatibilità degli approcci sperimentali appena illustrati con un paradigma enattivista del tipo che stiamo proponendo può essere chiamato in causa il sistema neurale comune a tutte le manifestazioni di *intentional attitude* fin qui citate, ovvero il sistema dei neuroni specchio (Rizzolatti et al. 1999). Infatti, i neuroni specchio, unità cerebrali nel cervello di alcuni mammiferi – umani compresi – che viene attivato sia quando si sta compiendo un'azione, sia quando si sta osservando qualcuno compiere la stessa azione, sono considerati fornire evidenze scientifiche a supporto di una visione incarnata ed enattivista della mente (Gallagher & Bower 2014; Paolucci 2020; Hutto 2008; Gallese et al. 2009; Sinigaglia 2009; Damiano 2010). Anche se in un primo momento ha ricevuto un'interpretazione prevalentemente in ottica simulazionista e della teoria della mente (Gallese e Goldman 1998, cfr. §3.3.4-§3.3.4.1), oggi il sistema dei neuroni specchio è ancora ritenuto essere una delle modalità attraverso cui la comprensione dell'altro avviene innanzitutto su un piano corporeo e senza ricorrere a ragionamenti inferenziali, ma in un'ottica compartecipata e condivisa, costituendo la base innata su cui costruiamo la nostra capacità interazionale. Come descritto da Damiano,

Il fatto che i *mirror neurons* non discriminino tra modalità motorie e modalità sensorie di specificazione di un'azione attribuisce a essi una proprietà funzionale singolare. Ne fa luoghi d'espressione di schemi motori i quali possono supportare sia l'attività di un soggetto che compie un'azione, sia quella di un soggetto che vi assiste. Ciò che i *mirror*

*neurons* realizzano naturalmente sono schemi d'azione *neutri* rispetto al ruolo svolto dal soggetto del movimento specificato. È una funzione che può essere indifferentemente esecutiva od osservativa, poiché il medesimo *pattern* neurale consente a un individuo di assumere sia una posizione di *prima*, sia una posizione di *terza persona* nell'attuazione dell'azione. [...] si tratta della contemporanea attivazione di uno stesso *pattern* neurale nell'osservatore e nell'esecutore di un'azione: la condivisione di uno schema di attività neuronale, che diviene *comune*. (Damiano 2010: 225)

Anche nel campo dell'indagine sull'interazione tra umani e robot, il ruolo dei neuroni specchio è stato spesso messo alla prova, producendo risultati che supportano la presenza dell'attivazione anche davanti ad artefatti umanoidi (Gazzola et al. 2007; Oberman et al. 2007)

In sostanza, gli studi che abbiamo presentato testimoniano del fatto che il robot può essere percepito a livello corporeo e situato «as a goal-oriented, predictable agent, able to evoke the same motor resonance as a human actor» (Sciutti et al. 2013: 343). Ma perché, allora, in molte indagini (Keijser et al. 2020; Stafford et al. 2014; Spatola 2019; Banks 2020; Fussell et al. 2008; Bartneck et al. 2007; Bartneck & Hu 2008, cfr. Thellman et al 2022) questo non si rispecchia nell'utilizzo di *propositional attitude* conformi? Cosa impedisce di assecondare la comprensione corporea che abbiamo del robot, quando arriviamo a dover descrivere verbalmente il suo comportamento?

In realtà, come abbiamo evidenziato all'inizio di questa indagine, ci sono molte evidenze del fatto che spesso è il linguaggio mentalistico – l'*intentional stance* – ad essere impiegato nella descrizione del comportamento del robot, ma questi provengono per lo più da studi riguardanti l'antropomorfismo (Zlotowski et al. 2015a; Epley et al. 2007; Fink 2012; Kim 2024; Fischer 2021), considerato uno dei fattori più cruciali da investigare per comprendere il processo di comprensione del robot come agente intenzionale (Epley et al. 2007; Spatola & Wykowska 2021; Spatola et al. 2021a; Marchesi et al. 2021a). Come analizzato nel capitolo precedente (§3.4.1), quindi, la cifra interpretativa che contraddistingue la maggior parte di questi studi si concentra sulla componente *mimetica* del robot, che include un aspetto di inganno: l'aspetto del robot ci conduce ad interpretare e descrivere il suo comportamento in termini mentalistici che però, a un'indagine più approfondita, non potrebbe essere supportata da correlati ontologici. In altre parole, quello che vedo è un tranello, una menzogna, realizzata dal robot per mezzo della sua somiglianza all'umano ma che non è sorretta dalla sua conformazione cognitiva interna. Il robot non ha una mente, come dichiarato da molti individui in risposta a questionari diretti (Banks 2020; Spatola 2019), tuttavia sono portato a descriverlo tramite linguaggio mentalistico assecondando la tendenza all'antropomorfizzazione. Un inganno percettivo, dunque, che ha come conseguenza la descrizione in termini mentalistici nonostante la consapevolezza della sua natura mendace. L'approccio enattivista ci ha consentito di riformulare questa visione, facendoci allontanare dalla considerazione del robot come ingannatore: il linguaggio antropomorfo non deve essere visto come la descrizione di uno stato di cose, ma come l'insieme di affermazioni

pragmatiche nel contesto interazionale: «Anthropomorphism is the recognition of ‘intersubjectivity’ in action» (Damiano & Dumouchel 2018: 6).

Ci troviamo di fronte a una situazione in cui, da una parte, una serie di studi, principalmente afferenti alle indagini sull’antropomorfismo, valida la presenza dell’attribuzione degli stati mentali per mezzo della presenza dell’antropomorfismo, mettendo alla prova i soggetti in compiti sperimentali che vogliono testare se il comportamento presenta tendenze a considerare, su più livelli, il robot come un essere umano, e evidenzia una generale tendenza ad assecondare tendenze antropomorfe; dall’altra, tramite la somministrazione di questionari diretti, l’indagine sperimentale sull’attribuzione dell’intenzionalità registra nella maggior parte dei casi una risposta negativa, con gli individui che propendono a negare esplicitamente stati mentali e ascrizioni mentalistiche ai robot. La differenza tra questi due tipi di risposta alle sollecitazioni procurate negli esperimenti risiede, secondo noi, nelle modalità di porre le domande: se il partecipante all’esperimento viene interrogato direttamente, con domande come «Do you think Ray [the robot] has a mind?» (Banks 2020: 406), la risposta è tendenzialmente negativa (2020: 412)<sup>155</sup>. Ma se il compito è predisposto in forme meno esplicite di verifica, come ad esempio nei report elaborati a seguito di un incontro con il robot (Fussell et al. 2008), allora ciò che prevale è la tendenza a usare un linguaggio mentalistico, plausibilmente assecondando l’*intentionality detection*.

La divisione tra *intentional* e *propositional attitude*, allora, si conferma un approccio utile alla HRI, che riesce ad illuminare aspetti che agli altri approcci all’intenzionalità restano invisibili. Infatti, gli approcci che abbiamo delineato (Haugeland 1990) concepiscono l’intenzionalità come un fenomeno la cui presenza può alternativamente essere o non essere riscontrata, attribuita o sviluppata, senza lasciare spazio a dinamiche di aggiustamento o di condizionamento da parte del contesto in cui avviene l’interazione. Adottando, d’altra parte, la prospettiva enattivista, possiamo fornire un’analisi secondo cui le forme corporee di comprensione predominano sulla considerazione del robot come agente interazionale potenzialmente capace di assecondare le dinamiche intersoggettive in contesti sperimentali in cui non c’è esplicita richiesta di un giudizio in tal senso, ma le *folk psychological narratives* che nutrono le nostre *propositional attitude* si configurano come esplicitamente orientate a negare la possibilità che il robot possa assumere il ruolo di un agente capace di interagire con noi al pari di un altro essere umano. In altre parole, quando ci viene richiesto di interpretare linguisticamente il comportamento del robot, le narrazioni che hanno strutturato la nostra capacità di interpretare il mondo tramite le *propositional attitude* ci spingono a rispondere negando al robot lo

---

<sup>155</sup> Un esperimento interessante (Spatola 2019) ha dimostrato che la risposta a domande del genere potrebbe essere influenzata anche dal concetto di ‘mente’ a cui il paziente aderisce: a una concezione dualista della mente corrisponderebbe una tendenza a negare la caratteristica al robot, mentre a una visione funzionalista si accoppia una maggiore apertura a considerare il robot, anche solo in futuro, una possibile entità dotata di mente.

statuto di agente intenzionale. Non è un caso che diversi studi evidenzino la discrepanza di assumere atteggiamenti intenzionali di fronte al robot tra partecipanti di cultura occidentale o asiatica. La variabilità culturale e socioeconomica, infatti, è uno dei fattori considerati influenzare in modo sostanziale la tendenza all'antropomorfizzazione e alla considerazione del robot come agente intenzionale (Marchesi & Wykowska 2024; Marchesi & Spatola 2024; Marchesi et al. 2021b, 2023; Spatola et al. 2021a, 2022) e diversi studi riportano una differenza circa la disponibilità ad attribuire stati mentali o intenzionalità al robot: popolazioni asiatiche, come cinesi e giapponesi, sono più inclini a descrivere il robot in termini antropomorfici rispetto a persone provenienti dagli Stati Uniti (Tan et al. 2018), dalla Germania (Bernotat & Eyssel 2018), dall'Australia (Haring et al. 2015), dall'Italia (Trovato & Eyssel 2017), ma anche da altri paesi considerati appartenenti alla sfera della cultura occidentale (Takahashi et al. 2016). Alcuni (Lim et al. 2021; Kodate et al. 2022; Yam et al. 2023) hanno ipotizzato che questo dipenda dal fatto che, nelle narrazioni che hanno circolato all'interno delle due differenti matrici culturali, la figura del robot è dipinta in modi estremamente diversi<sup>156</sup>:

- da una parte, nella cultura occidentale prevale una dimensione antierica del robot, sulla scia di quella paura ancestrale verso le creature artificiali che abbiamo descritto nel capitolo I, che ha contribuito a costruire la figura del robot come minaccia per l'umanità – da Frankenstein a Čapek, dal Golem ai racconti di Philip K. Dick – e che prevale tuttora nella fantascienza di settore soprattutto nel cinema, con film come *Terminator* (1984), *Ex Machina* (2014) o la serie televisiva *Westworld* (2016-2022). Inoltre, lo sviluppo di un dibattito sull'intelligenza artificiale in generale, anche questo al centro di numerose narrazioni diffuse in diversi ambiti della cultura popolare – come, ad esempio, il film *Her* (2013) –, potrebbe contribuire sostanzialmente a far propendere le persone, su esplicita richiesta, ad affermare che il robot non abbia capacità interazionali, non sia di fatto davvero intelligente, ma abbia la capacità di ingannare il suo interlocutore per fargli credere il contrario.
- dall'altra, la cultura giapponese, ad esempio, ha fornito un nutrito campione di narrazioni, dai *manga* agli *anime* – come, ad esempio, *Astro Boy* (1963, 2009), *Power Rangers* (1993-1995), *Doraemon* (1969-1996), *Neon Genesis Evangelion* (1995-1996) – in cui il robot è visto come una figura amichevole, spesso aiutante dell'umano. Inoltre, soprattutto in Giappone la diffusione dei robot in contesti quotidiani, dall'hotel al supermercato, dai ristoranti agli ospedali, che affiancano stabilmente gli esseri umani nello svolgimento di lavori e mansioni anche molto diversi fra di loro (cfr. Sone 2016) contribuisce senza dubbio a costruire un

---

<sup>156</sup> Il ruolo delle narrazioni è spesso messo in risalto rispetto all'appartenenza culturale dei soggetti intervistati: come suggerito da Lim et al. (2021), la globalizzazione delle culture, e di conseguenza il contatto con narrazioni anche molto diverse rispetto a quelle della cultura di arrivo, ha contribuito alla diffusione interculturale di prodotti come film, racconti, romanzi, fumetti e cartoni animati, attenuando in parte le differenze.

immaginario in cui agenti artificiali possono effettivamente portare avanti compiti in cui c'è bisogno di interagire in modo dinamico, costruendo un rapporto di fiducia più stabile rispetto ad altri contesti in cui prevalgono narrazioni apocalittiche.

Infine, se consideriamo il fenomeno attraverso la distinzione tra *propositional* e *intentional attitude*, vediamo che la possibilità che esperimenti volti ad indagare come e se il robot sia individuato in quanto agente intenzionale non comporta contraddizioni a livello cognitivo. In sostanza, veicolando l'*intentional* e le *propositional attitude* due tipi diversi di intenzionalità anche se intrecciate, e non due manifestazioni dello stesso meccanismo sottostante, i modi espliciti e impliciti di valutarli potrebbero effettivamente produrre risultati diversi tra loro, senza per questo implicare dissonanze cognitive (*infra* § 5.2.2). Il connubio tra approccio neo-comportamentista e neo-pragmatista all'intenzionalità, raccolto sotto i principi della visione enattivista, ci consente di vedere la questione dell'*intentionality detection* come un processo di aggiustamento su più livelli, che si configura contemporaneamente allo sviluppo dell'interazione e che mette in gioco processi dinamici relativi alla sfera corporea e contestuale, individuale e collettiva, astratta e pragmatica, ognuna delle quali risente degli sviluppi che avvengono nelle altre sfere in cui si realizza la nostra capacità di dare senso alle azioni degli altri. È indagando i) il modo in cui siamo evoluti a livello corporeo e cognitivo e ii) il modo in cui siamo entrati a far parte di una comunità linguistica e normativa fatta di pratiche sociali, narrazioni e tecnologie che possiamo comprendere come il modello di cognizione che utilizziamo per indagare i processi cognitivi in ambito intersoggettivo tra umani può essere utile anche per comprendere cosa succede nella HRI.

La questione dell'intenzionalità in HRI, quindi, potrebbe sembrare chiarita. La comprensione del robot come agente intenzionale passa dalla individuazione, dal rilevamento, dal riconoscimento di una matrice intenzionale nei suoi movimenti, fattore che enattiva movimenti e comportamenti a loro volta intenzionali in chi guarda e garantisce la possibilità della mutua comprensione. Tuttavia, non abbiamo ancora affrontato la questione dell'intenzionalità genuina per l'approccio enattivista, che a uno sguardo esperto sicuramente non sarà sfuggita<sup>157</sup>.

L'intenzionalità enattivista è strettamente legata alla presenza di *sense-making*, ovvero quella capacità dell'organismo biologico di costruire senso a partire dalla sua struttura autopoietica e dall'autonomia che questa produce (Maturana & Varela 1986; Varela et al. 1991; Thompson 2007). La comprensione reciproca, inoltre, che si verifica in interazione tra due sistemi autonomi e autopoietici si appoggia alla nozione di intercorporeità (Merleau-Ponty 1945; Varela et al. 1991;

---

<sup>157</sup> Le prime fasi di realizzazione della necessità di riorganizzare in parte la concezione enattivista dell'intenzionalità in relazione agli studi di HRI sono avvenute durante una conversazione con Tom Froese, alla 4<sup>o</sup> edizione della "International HRI Summer School" tenutasi in Polonia, nel settembre 2023.

Gallagher 2017, 2021), che prevede l'incontro di due corpi vivi, i quali si coordinano a vicenda sulla base di una spontanea comprensione incarnata e reciproca realizzata grazie alla qualità situata e corporea dell'esperienza di ogni individuo che si trova a interagire. Come è possibile, allora, che i robot esibiscano un qualche tipo di *intentional attitude*, dal momento che non sono individuabili né come organismi autopoietici, né come agenti la cui autonomia è sufficiente per produrre capacità di *sense-making* al pari di quelle di altri organismi viventi, né come corpi vivi capaci di esperire ed esprimere l'intercorporeità? C'è bisogno di ritornare sui presupposti teorici sui cui si costruisce l'approccio enattivista all'intenzionalità nel dominio dell'interazione sociale, al fine di capire come sia possibile fornire una descrizione dell'*intentionality detection* nella HRI senza tradire i presupposti incarnati intrinseci all'impianto teorico che abbiamo delineato.

#### **4.4 La chiave dell'intenzionalità: il primato dell'*engagement***

Il problema che si pone in questo momento a un approccio che si vuole enattivista e che deve rendere conto dei processi di comprensione che avvengono tra robot e umani nei termini dell'intenzionalità è relativo al rapporto tra questa e il processo di *sense-making*.

Chiariamo meglio la questione. Le posizioni in merito all'origine dell'intenzionalità, o intenzionalità originale, degli approcci che abbiamo considerato grazie a Haugeland (1990) sono molto diverse fra loro, ma l'ibridazione che abbiamo proposto come base per considerare l'intenzionalità attraverso lo sguardo enattivista condivide la natura externalista: da una parte, l'approccio neo-comportamentista di Dennett pensa all'intenzionalità come attribuita da un agente attraverso l'adozione dell'*intentional stance*; dall'altra, l'approccio neo-pragmatista la concepisce come una proprietà derivante dalle pratiche situate e collettive in cui l'individuo si trova ad agire. Il punto è, però, che tali considerazioni sull'origine dell'intenzionalità valgono, se trasferite all'approccio enattivista delle *attitude* che abbiamo preso dalla REC, solamente per le *propositional attitude*: l'intenzionalità caratterizza la descrizione del comportamento agli occhi di chi guarda compiere l'azione (neocomportamentismo) e la possibilità di esercitare tale *expertise* si crea entrando a far parte di un contesto narrativo e pratico (neo-pragmatismo). Tuttavia, se vogliamo rispettare l'assunto secondo cui la modalità di comprensione realizzata tramite le *propositional attitude* abbia una struttura diversa rispetto a quella corporea più immediata e automatica delle *intentional attitude*, allora la loro origine non può essere la stessa. L'intenzionalità corporea, per come dimostrata nelle *intentional attitude*, deve avere a che fare con la natura dell'essere umano in quanto organismo situato nel suo ambiente, intrecciato alle dinamiche di accoppiamento che si instaurano tra corpo, cervello e mondo.

Abbiamo già incontrato alcune forme di questa intenzionalità enattivista corporea, intrinseca, che garantiscono sia la possibilità delle *intentional attitude* sia, su piani più complessi ed evoluti, di *propositional attitude*, ovvero la *Ur-intentionality* (*infra* §4.3.1; Hutto & Satne 2015; Hutto & Myin 2017). In realtà, anche altri esponenti della corrente enattivista hanno preso in considerazione questo aspetto dell'intenzionalità, insieme alla necessità di individuare la sua struttura caratteristica e le pertinenze alle quali si riferisce, elaborando terminologie differenti ma che possono tutte essere considerate riferirsi allo stesso concetto. *Skilled intentionality* (Kiverstein & Rietveld 2015), *enactive intentionality* (Gallagher 2017), *neo-pragmatic intentionality* (Gallagher & Miyahara 2012), *autopoietic intentionality* (Varela 1992a), *motor intentionality* (Dreyfus 2002), *biological intentionality* (Thompson 2007), *proto-intentionality* (Varela et al. 1991: 68), *somatic intentionality* (Sachs 2014): tutte queste accezioni indicano una forma basilare di intenzionalità che afferisce alle dinamiche corporee di comprensione e interpretazione del mondo e che funziona da *ground* per la stratificazione di tipi più complessi, anche linguistici, di intenzionalità. Come sottolineato da Steiner (2025: 7, corsivi dell'autore), per tutte queste declinazioni, «intentionality is *intrinsically* real: it does not depend on some attributions by an observer for being had and for exercising itself; and it does not depend on the inclusion of the creature in linguistic, social and normative practices. What intentionality *does* is important».

Cosa *fa*, allora, l'intenzionalità corporea? Rimanendo ancorati alle definizioni della REC, ad esempio, la *Ur-intentionality* «explain how organisms relate to and connect with targeted aspects of their worldly environments» (Hutto & Myin 2017: 50). In questo senso, l'intenzionalità si configura come il tramite attraverso cui l'organismo e l'ambiente vengono messi in relazione, la proprietà della cognizione orientata all'azione di generare significato per l'organismo. In un modo diverso, Varela afferma la stessa visione, legando strettamente la *situatedness* dell'organismo alla sua capacità di dare senso all'ambiente in cui si muove:

[...] the emergent, parallel, and distributed dynamics is inseparable from the *constitution of a world*, which is none other than the surplus of meaning and intentions carried by situated behavior. If the links to the physical environment are inevitable, the uniqueness of the cognitive self is this constant genesis of meaning. Or, again to invert the description, the uniqueness of the cognitive self is this constitutive *lack* of signification which must be supplied faced with the permanent perturbations and breakdowns of the ongoing perceptuo-motor life. Cognition is action about what is *missing*, filling the fault from the perspective of a cognitive self. This view amounts to a biology of intentionality. (Varela 1992b: 13)

Questa visione lega *sense-making* e intenzionalità in modo strettissimo, intrecciando le capacità dell'organismo di dare senso al suo ambiente alle qualità che l'esperienza del senso mostra proprio nel momento del suo farsi. Inoltre, *sense-making* e intenzionalità sono intrecciate anche alla presenza della vita, intesa come attività cognitiva realizzata per mezzo della struttura incarnata del sistema.



Questo concetto è perfettamente espresso da Thompson, il quale identifica l'intenzionalità con le capacità, anche minime, dell'organismo di dare senso al suo ambiente – «sense-making [...] is none other than intentionality in its minimal and original biological form» (2007: 147); «Intentionality arises from the operational closure and interactive dynamics of autopoiesis» (2007: 159); «intentionality first emerges in nature in the form of autopoiesis and sense-making» (Ibidem); «original intentionality [is] understood as the sense-making capacity proper to autopoietic units» (2007: 453) – ma anche con la proprietà del corpo vivo - «intentionality is a fundamental characteristic of the lived body» (2007: 15).

Ora, questa visione ha una profonda implicazione per l'impiego del concetto di intenzionalità enattivista negli studi di HRI. Seguendo strettamente i principi enucleati fin qui, infatti, ciò significa che non possa esserci intenzionalità corporea, espressa o percepita, per quegli agenti che non possiedano capacità di *sense-making*, che non siano autonomi nel senso biologico del termine e che non ancorano la loro attività cognitiva a schemi sensorimotori di realizzazione della propria identità: come sottolineato anche da Di Paolo (2003), i robot restano esclusi. Questa visione, però, risulta in contrasto con gli esperimenti sullo sguardo anticipatorio e sulla risonanza motoria, che individuano nelle risposte interazionali dell'umano una calibrazione del proprio comportamento a seguito dell'identificazione del robot come agente intenzionale. Come è possibile?

Affrontando le varie declinazioni contemporanee della nozione di intenzionalità nel dominio enattivista, abbiamo sempre sottolineato che questa è vista come una proprietà dell'intero organismo, che si esprime nelle azioni che lo stesso compie nell'ambiente in cui è inserito. Tuttavia, sembra che, in alcuni passaggi, questa visione abbia lasciato spazio a una meno distribuita, in cui l'intenzionalità non è più *tra* l'organismo e l'ambiente, ma è «'visible' in the surface flow of agents' motions» (Ansuini et al. 2014). Probabilmente in relazione alla rivalutazione delle istanze fenomenologiche, che hanno da sempre – ma sempre più – caratterizzato la visione enattivista almeno nelle concettualizzazioni più popolari dell'intenzionalità (Gallagher 2017; Thompson 2007), questa ha intrapreso un percorso verso la sua localizzazione *sul* corpo dell'agente, espressa, sì, nelle azioni svolte da questo ma strettamente legata alle qualità della struttura che rende un corpo vivo tale. È chiaro che, per descrivere il processo di cognizione nell'organismo umano o la struttura della sua esperienza incarnata, non ci sia la stretta necessità di tenere distinte le due nozioni, dal momento che si manifestano in concomitanza e non possono che essere viste come due facce della stessa medaglia: la capacità di produrre senso e la sua direzionalità relazionale con aspetti o materialità dell'ambiente si manifestano simultaneamente nelle interazioni incarnate rivolte ad *affordance* e negli schemi corporei che regolano il comportamento. Ma, come testimoniato dalla serie di esperimenti che rilevano la presenza di *intentional attitude* sia nel comportamento del robot, sia negli *attunement*

dell'umano che si trova a interagire con lui, nel campo degli HRI questa aderenza dell'intenzionalità al corpo dell'agente potrebbe risultare problematica, dal momento che non si riuscirebbe a rendere conto dei motivi che portano a considerare il robot un agente intenzionale in interazione.

La soluzione che proponiamo per ovviare a questo *impasse* è quella di recuperare le radici della concezione enattivista dell'intenzionalità, illuminando un altro senso in cui intenzionalità e *sense-making* sono correlati, a partire dalla considerazione di Francisco Varela secondo cui l'intenzionalità è il modo attraverso cui il *sense-making* si dispiega nel comportamento (1992a, 1992b). In questo senso, l'intenzionalità non è localizzata o attribuita semplicemente al corpo, ma emerge dalla rete interdipendente di relazioni che si stabiliscono nell'interazione dinamica tra l'agente e il suo mondo.

In realtà, le nozioni non sono adottate in modo mutuamente esclusivo, e anzi vanno spesso di pari passo, come ad esempio succede in Gallagher:

Intentionality is determined by what the agent is doing and what the agent is ready to do – i.e., in the agent's sensorimotor skills to cope with the situation at hand – and that holds for both stepping off a curb and stepping on the brake, and for any interaction that might follow. [...] On the enactive view, one doesn't need to go to the level of mental states [...] to encounter intentionality – operative intentionality is in the *movement*, in the *action*, in the *environmentally attuned responses*. This operative intentionality is (non-derived, primary) intentionality. (Gallagher & Miyahara 2012: 16)

Ciò che si vuole fare, allora, è riprendere l'idea di intenzionalità nella sua versione di qualità esibita dall'*engagement* dell'organismo con l'ambiente, ovvero espressa non solamente nell'azione performata sulla superficie dei movimenti dell'agente, ma come proprietà della relazione che si istituisce *tra* agente e ambiente, qualità del movimento che, nel suo svolgersi, allaccia legami materiali e agentivi enattivando gli *affordance* materiali e sociali offerti dall'ambiente e dal contesto in cui vengono realizzati. Per cercare di capire se questa ipotesi è in accordo con i presupposti enattivisti, dobbiamo spostare l'attenzione sul rapporto dinamico istituito a livello corporeo tra organismo e ambiente, ovvero il legame tra l'intenzionalità e l'*agency*.

#### 4.4.1 «*Intentions in every case come already clothed in agency*»: l'intenzionalità enattiva agli occhi di chi guarda

Nelle scienze cognitive e nella filosofia della mente, il legame tra *agency* e intenzionalità è stato da sempre stato considerato molto stretto. Infatti, l'*agency*, o agentività, è considerata classicamente la capacità dell'agente di compiere azioni rivolte verso il mondo e implica il concetto di intenzionalità dal momento che

a being has the capacity to exercise agency just in case it has the capacity to act intentionally, and the exercise of agency consists in the performance of intentional actions and, in many cases, in the performance of unintentional actions (that derive from the performance of intentional actions [...]). (Schlosser 2019)

Fino all'avvento delle teorie *embodied* riguardo la mente, questi due aspetti della cognizione sono stati considerati in un legame causale diretto: come abbiamo visto quando abbiamo descritto l'approccio neo-cartesiano, secondo le teorie classiche delle scienze cognitive basate sulla rappresentazione degli stati mentali e sui principi computazionalisti, è l'organizzazione funzionale dell'architettura cognitiva che comporta che l'agente abbia la capacità di agire intenzionalmente, ed è la presenza di certi stati mentali a rendere possibile l'azione specifica e l'agentività del sistema cognitivo. In altre parole, per gli approcci classici, l'intenzionalità causa l'*agency*, secondo un approccio definito anche *in-the-haed*, «which emphasize computational processes in the brain that produce and consume representations, such as facts about my intentions and the actions I perform» (Buhrmann & Di Paolo 2017: 212).

Con la riformulazione offerta dall'approccio enattivista, tuttavia, il rapporto tra agentività e intenzionalità del sistema si ridefinisce, anche e soprattutto in rapporto alla cognizione: quest'ultima, infatti, è innanzitutto una forma di *embodied action* (Varela et al. 1991: 172-179). In particolare, per l'enattivismo l'azione diventa uno degli elementi fondativi della cognizione, il mezzo attraverso cui il sistema autopoietico e autonomo abita il mondo all'interno del quale è inserito e realizza (*bring forth*) l'ambiente tramite le possibilità di intervento e di modifica consentite dal suo *embodiment* (Varela et al. 1991; cfr. §2.3.1.2). In questo senso, l'agentività è elemento costitutivo delle capacità intenzionali del sistema cognitivo nel suo complesso (Di Paolo et al. 2017; Bruinenberg 2018; Segundo-Ortin 2020), e si fonde all'intenzionalità non perché questa è causa o motore delle azioni dell'organismo, ma perché si realizza come capacità inerente alle azioni performati dall'agente di essere circa qualcosa, rivolte a un aspetto o proprietà dell'ambiente, relative a un processo di mantenimento dell'organizzazione interna come vitale e strettamente legate al complesso di valori sulla base del quale l'organismo può continuare a sopravvivere come sistema cognitivo (Di Paolo et al. 2017). Tale aderenza costitutiva è garantita dalla chiusura operativa del sistema, che allo stesso tempo rimane aperto agli scambi con l'ambiente, mantenendo quell'equilibrio di autonomia che è fondazione della capacità di aprirsi a un mondo e di costruire un ambiente significativo in termini di azioni pragmatiche e incarnate:

Cognition or sense-making is the intentional and normative engagement of the system with its environment. One of the basic propositions of the enactive approach is that being autonomous is a necessary condition for a system to embody original intentionality and normativity. Unless the processes that make up a system constitute that system as adaptive self-sustaining unity, there is no perspective or reference point for sense-making and hence no cognizing agent. Without autonomy (operational closure) there is no original meaning; there is only the derivative meaning attributed to certain processes by an outside observer (Thompson & Stapleton 2009: 6)

Dunque, l'enattivismo ci invita a considerare l'intenzionalità, e quindi la intrinseca significatività dell'azione per l'agente, non come un tratto mentale statico, ma come un processo dinamico e

relazionale che si dispiega nel tempo e nello spazio e radicato nell'interazione con l'ambiente. In questo contesto, l'intenzionalità non può essere separata dall'azione e dalla capacità dell'organismo di performarne. Anzi, l'unico modo attraverso cui è possibile per l'organismo realizzare forme di intenzionalità è tramite l'azione incarnata all'interno di un contesto fisico, materiale e corporeo: gli organismi cognitivi esibiscono *agency* quando interagiscono con il mondo attraverso comportamenti auto-regolati e orientati verso determinati obiettivi.

Ora, se traduciamo questo discorso nei termini della cognizione sociale, avremo un approccio che vede la comprensione dell'altro avvenire in un processo di continua generazione di significato, realizzato nel corso del mantenimento dell'autonomia del sistema tramite azioni volte alla sussistenza dei cicli sensomotori e delle possibilità di accoppiamento con l'ambiente, in cui viene allo stesso tempo implicato un altro agente avente le stesse capacità (De Jaegher & Di Paolo 2008; Fuchs and De Jaegher 2009; De Jaegher 2009; Gallagher 2021; Froese & Gallagher 2012). Questo processo, in cui il *sense-making* è realizzato attenzionando una porzione dell'ambiente che comprende l'agente e le sue capacità di azione al suo interno, oltre che dalle sue possibilità di interazione dialogante con l'altro, è supportato da una forma di percezione attiva che si costituisce come un'azione incarnata (Varela et al. 1991): nel dominio interazionale intersoggettivo, percepire l'altro significa enattivare l'insieme delle azioni che potrebbero essere intraprese come significative se concatenate al suo comportamento<sup>158</sup>.

Abbiamo visto che il processo di comprensione dell'altro avviene proprio in relazione alla individuazione percettiva di una relazione intenzionale, la quale si dispiega nel comportamento dell'agente che diventa immediatamente significativo per l'altro, e a cui egli si concatena nella danza pragmatica dell'interazione sociale. Accade che, nell'incontro con un altro agente intenzionale, alcuni accoppiamenti agentivi – con porzioni del proprio ambiente e con l'interattante – che questo performa sono individuati come immediatamente significativi, proprio perché nell'interazione si costituiscono come contingenze sociali che significano nel loro modo di svilupparsi in quanto performate da un agente intenzionale. Come illustrato da McGann & De Jaegher nel caso della cognizione infantile,

Objects in the infant's world either react to his motions with high contingency (moving immediately after a movement of the infant, for instance an object that is kicked by the infant) or with a low contingency (moving at a slight delay, as does another person). [...] On the enactive, interactive account of perception, contingencies [...] characterize an individual's *interaction* with a certain stimulus. As such, social contingencies are in the

---

<sup>158</sup> «The enactive approach to understanding the mind means that cognition is never conceived of as a series of progressive steps beginning when the sensory surfaces are stimulated and culminating in the triggering of peripheral muscle movements. Perception is a facet of sense-making rather than a distinct “input” phase. It must be understood within the context of valued, skillful activity. Perception is not structured by the character of bodily sensitivities alone [...] but also by the skills that the agent is enacting, deploying at a given time. [...] If perception is the skillful engagement with the physical world, social perception is the skillful handling of social situation». (McGann & De Jaegher 2009: 8; 13)

first place the processes that individuals themselves participate in and construct. Social contingencies are features of social interactions, not of objects perceived from a distance in order to then help categorise the object, e.g. as living, social or static. (Mc Gann & De Jaegher 2009: 15)

Quello che vogliamo sottolineare è che la necessaria e costitutiva espletazione dell'azione affinché possa individuarsi l'intenzionalità intrinseca ai processi di *sense-making* messi in pratica dall'organismo ci fa spostare l'attenzione dalle proprietà intrinseche alla cognizione dell'agente alla sua manifestazione agentiva: è attraverso le azioni nell'ambiente e la costituzione di un comportamento dell'agente che, in interazione, è possibile individuare la capacità intenzionale di dare senso al mondo.

One's motor intentions (M-intentions) and intentions-in-action (P-intentions)<sup>159</sup> are just that – intrinsic to the motor kinematics and action itself. M- and P- intentions reflect the fact that the experiencing agent is intentionally engaged with the world through actions and projects that are not reducible to simple internal mental states, but involve an intentionality that is motoric and bodily. Actions have the intentionality because they are directed at some goal or project, and this is something we can see in the actions of others. (Gallagher 2021: 123)

In sostanza, come abbiamo visto, «intentions shape action kinematics» (Gallagher 2021: 123), ed è afferrando la cinematica dell'azione che riusciamo a individuare l'intenzionalità, in un processo che è basato su sistemi immediati di riposta motoria rintracciabili nei pattern comportamentali incarnati – come lo sguardo anticipatorio – e correlata alla capacità di individuare il movimento intenzionale tramite la sua risonanza nella nostra struttura neuro-cognitiva – *motor resonance*. Gli stati intenzionali sono, sì, stati corporei, ma in ottica enattivista sono meglio compresi come stati delle interazioni tra organismo e ambiente (Steiner 2023: 11; 2025): in un contesto in cui ci sono due organismi che si trovano a interagire (organismo1 e organismo2), è nella messa a sistema di i) le capacità percettive date dalla significatività del mondo per l'organismo1; ii) lo svolgersi dell'azione nello spazio e nel tempo realizzata dall'organismo2; iii) il concatenamento delle possibilità di azione e delle loro effettive concretizzazioni in dialogo; iv) e il fatto che questo avvenga reciprocamente e contemporaneamente che si crea quella nebulosa relazionale di scambi mutuali e co-costruzioni di senso identificabile con l'interazione. Organismo1, per riuscire a partecipare nella costruzione di senso resa possibile dall'incontro con organismo2, deve riuscire a individuare gli appigli a cui

---

<sup>159</sup> Quelle che Gallagher (2021: 48; 54-55) definisce *M-intentions* e *P-intentions* possono essere considerate anche due specificazioni della nozione di *intentional attitude*. Le *M-intentions* coinvolgono processi di controllo che regolano lo svolgimento dell'azione secondo una normatività motoria; le *P-intentions* specificano l'azione nei termini particolari dei movimenti richiesti nel suo stesso svolgersi, si aggiustano o si creano in relazione a un compito motorio – identificabili con la nozione di *intention-in-action* data da Searle (1983). Entrambe, dunque, riguardano un tipo di intenzionalità incarnata che non ha a che fare con contenuti proposizionali, ma solamente con azioni situate dell'organismo nell'ambiente. A queste due, Gallagher aggiunge anche le *D-intentions*, identificabili invece come una declinazione delle *propositional attitude*: «D-intentions relate to prior deliberation processes that allow us to formulate pur relatively long-term (future) goals» (Gallagher 2021: 47). Si può leggere, quindi, la citazione sostituendo i termini usati da Gallagher con quelli proposti dalla REC senza pericolo di fraintendimenti.

agganciarsi nel flusso dell'azione, deve afferrare il significato complessivo per ricavarne la possibilità di inserirsi nel flusso costante della produzione e interpretazione di senso: ciò è possibile attraverso l'individuazione della struttura agentiva, ovvero dell'*engagement* realizzato dall'organismo per mezzo della sua capacità di generare la sua attività nell'ambiente (*interactional asymmetry*, Di Paolo et al. 2017; Barandarian et al. 2009), la quale porta nella sua struttura cinematiche l'impronta delle intenzioni che le aderiscono.

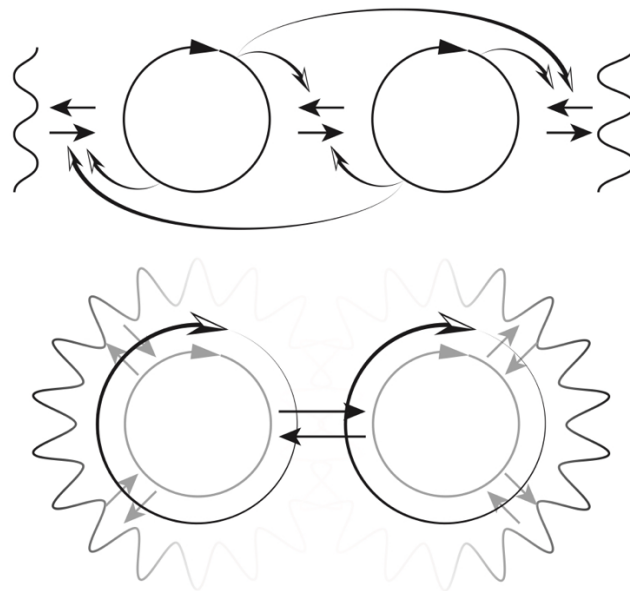


Fig. 4.4. Schematizzazione di una interazione sociale<sup>160</sup> (Di Paolo et al. 2018: 68).

Di Paolo et al. (2018) descrivono in termini precisi questo tipo di incontro tra agenti sensorimotori incarnati, ovvero sistemi autonomi che costruiscono e intraprendono interazioni sociali, le quali acquistano una propria autonomia, una chiusura operativa che garantisce all'interazione di prendere forma e indirizzare le agentività – e quindi le intenzionalità – di ognuno degli agenti coinvolti. È questa struttura autonoma dell'interazione che, in modo ricorsivo, indirizza e rimodula le agentività, le intenzionalità (sia come processi attentivi, sia come capacità di significare) e quindi il *sense-making* degli agenti coinvolti: «It is possible for individual sense-making to be *affected* by the activity of others without this influence being a regulatory one [for the engagement between agents]» (Di Paolo et al. 2018: 146; *infra* §5.2). L'*agency* è lo strumento fondamentale attraverso cui i sistemi si co-regolano e si comprendono a livello motorio, costruendo incassamenti di azioni rivolte verso l'altro e verso l'ambiente che condividono in cui l'intenzione si annuncia come sua intrinseca

<sup>160</sup> Descrizione della figura: «Social interaction. Two schematic depictions of coregulated activity during the interaction between two autonomous agents. Top: Coregulation of couplings in basic agents. Horizontal arrows indicate the coupling between each agent and its environment (wavy lines) and between them. The curved arrows represent the regulation each agent makes on its own couplings and on the other's. Bottom: Coregulated coupling in sensorimotor agents. The straight arrows again represent couplings with the environment and between the agents, but now the closure of the regulatory curved arrows is represented explicitly». (Di Paolo et al. 2018: 68)

caratteristica. In altre parole, «intentions in every case come already clothed in agency» (Gallagher 2021: 56): le intenzioni del tipo motorio, che abbiamo visto garantire una comprensione interindividuale a livello implicito senza bisogno di ragionamenti inferenziali o strutture linguistiche complesse a sorreggerle, si manifestano sempre attraverso l'intervento agentivo.

Ora, queste riflessioni, se applicate al contesto della HRI, possono farci procedere nella considerazione dell'intenzionalità almeno parzialmente: possiamo senza dubbio ammettere che l'intenzionalità non si dia come una proprietà nuda della mente interna all'agente, ma come un processo di sviluppo agentivo che si dispiega e prende forma solo attraverso le sue azioni. In questo modo, ciò che viene sottolineato è la possibilità di interpretare l'intenzionalità non solamente come una proprietà del singolo individuo o della sua capacità cognitiva, ma come una caratteristica distribuita del suo comportamento. Ciò concede di vedere il processo di *intentionality detection* non come la rilevazione di una caratteristica individuale, messa in campo dalla capacità cognitive del singolo, ma come l'individuazione di un'intenzionalità delle azioni il cui riconoscimento porta a possibilità di concatenamenti interazionali: il robot non è individualmente possessore di un'intenzionalità genuina, ma avviluppato nelle dinamiche intenzionali dell'interazione sociale che gli conferiscono abilità pragmatiche di co-regolazione reciproca. Gli esperimenti che abbiamo chiamato in causa per dimostrare la possibilità che il robot sia compreso a livello corporeo come un agente dalle capacità intenzionali supportano la possibilità di intendere l'intenzionalità non solamente in termini individualistici, ma in quanto capacità espressa nell'*engagement* dell'organismo nel suo ambiente; non confinata nel corpo dell'agente, ma diffusa nell'arco spaziale delle possibilità di azione disegnato dalle sue forme di *embodiment*, in una commistione di materiale, intenzionale e sensorimotorio che tiene insieme corpo e mondo, cognizione e percezione, autonomia e *sense-making*, dialogo e individualità: l'intenzionalità può così diventare un aspetto enattivato tramite la rilevazione di *perceptual contingencies* intrinseche all'interazione sociale. Come sottolineato da De Jaegher (2009), il processo di percezione veloce, significativa e diretta che soggiace alla possibilità di capire a livello corporeo l'altro in interazione non deve essere visto come un meccanismo che va solamente dall'agente al mondo, ma come un complesso di influenze su più livelli che includono l'interazione come terreno e fattore cruciale affinché l'individuo possa effettivamente esercitare le proprie capacità di *sense-making*:

If movements and sense-making are foundationally linked, as seen above, and people's movements can be interindividually coordinated to such an extent that patterns of coordination may take on a 'life of their own', we have a way of conceiving of a sense-making that is truly intersubjective. When movements fundamentally intertwined with sense-making can interindividually coordinate, this means that *sense-making activities can coordinate interindividually*. This opens up a range of socially cognitive phenomena. At the least, interpersonal movement coordination can influence individual sense-making. [...] More intricate possibilities are also opened up: we can now think about how intentions

can be transformed – even generated – in social interaction [...]. In other words, we are no longer tied to thinking of intentions as individually formed and then brought into a social situation to be interpreted, simulated, or directly perceived. Instead, we now have the opportunity to gain an insight into *intersubjectivity* in the most literal sense of the word. (De Jaegher 2009: 539-540)

Ciò ovviamente non ha dirette conseguenze sulla considerazione dell'intenzionalità originale per l'enattivismo. Non si vuole affermare, infatti, che sia possibile avere forme di intenzionalità genuina, ovvero legate alla sopravvivenza e al mantenimento dell'autonomia del sistema, che non si poggino su strutture non autonome o non autopoietiche. Il discrimine tra forme di intelligenza e cognizione come quella umana, ovvero del tipo organico, e forme di intelligenza sintetica sembra ancora essere l'autopoiesi e la capacità di *sense-making* che questo tipo di struttura organizzazionale della materia produce nell'organismo biologico (cfr. §2.3.1.2).

In contrast to organisms, the robots of today are “agents” in an impoverished sense of the word. Even when embedded in a sensorimotor loop of situated interaction their behavior can be fully described simply as movement as opposed to action. Their “world” is only a set of external variables that may affect such moving trajectories. As worlds go, a robot's quite devoid significance in that there is no sense other than the figurative in which we can say that a robot *cares* about what is doing. A robot may show cleverness of design and some degree of adaptivity, but this by itself still puts in not in the same class as animals but closer to the thermostat. (Di Paolo 2003: 10)

Il robot è costitutivamente diverso dagli organismi biologici dal momento che non condivide il valore della *sopravvivenza* come «the mother-value of all values» (Di Paolo 2003: 8-9): il robot non è vivo, non deve adattarsi forzatamente alle necessità impartite dal suo proprio ambiente affinché la sua organizzazione interna non si disgreghi, non deve mantenersi vitale attraverso scambi costanti con il proprio ambiente che mettono costantemente a repentaglio la sua capacità di realizzarsi in quanto individuo. La precarietà dell'equilibrio che porta alla nostra forma di vita è proprio ciò che ci consente di sviluppare genuina capacità di generazione del significato e intenzionalità non-derivata, ed è quella che manca ancora al robot affinché possa sviluppare una forma genuina di intenzionalità. Tuttavia, i robot sono in qualche senso agenti, «the external behavior of the robot resembles that of an agent whose goal is internally generated» (Di Paolo 2003: 16), e questo ha delle conseguenze sull'interazione con l'essere umano, sulla possibilità che questa si faccia sociale, sul portato agentivo che il robot offre all'interazione stessa e sull'effetto che fa all'essere umano che si trova alla sua presenza. Quindi, se non attraverso l'implementazione degli stessi processi autopoietici che portano alla strutturazione del *sense-making* organico, come riescono i robot a manifestare queste forme – anche ridotte – di agentività, comunque abbastanza efficaci per enattivare risposte da parte dell'umano?



Come prospettato da Di Paolo, una soluzione percorribile in questo senso è relativa alla possibilità di implementare meccanismi attraverso cui generare *forme* di cognizione che non si basino sulle stesse strutture biologiche, ma che ne sintetizzino il funzionamento:

We may invest our robots not with *life*, but with the mechanisms for acquiring *a way of life*, that is, with habits. This may be enough for them to generate a natural intentionality, not based now on metabolism, but on conservation of ‘one’ way of life as opposed to ‘another one’. How this sort of intentionality relates to the intrinsic generation of intentions by a metabolizing system remains an open issue that need further investigation. (Di Paolo 2003: 13-14)

Nell’essere umano e, per quanto permette il concetto di autopoiesi, in tutte le forme di vita biologiche, il *sense-making*, l’intenzionalità e l’*agency* si presentano insieme, dal momento che la seconda è una diretta proprietà del primo, che si realizza nelle forme dell’ultima: non c’è *sense-making* senza intenzionalità, visto che questa è una sua qualità intrinseca; non c’è intenzionalità senza agentività, la quale ne permette la realizzazione. Ma è possibile riuscire a *sintetizzare* forme di intenzionalità che *mimino* il modo attraverso cui, solitamente, negli agenti intenzionali si esprime la direzionalità del *sense-making*? In altre parole, possiamo rintracciare l’*intentionality detection* nell’interazione tra umani e robot analizzando il modo in cui il robot dispiega le proprie azioni nell’ambiente che condividiamo?

#### 4.4.2 Il ritmo del vivente: evidenze dall’implementazione del movimento biologico

Alcune conferme riguardo alla possibilità che l’*intentionality detection* possa essere nutrita dalla capacità del robot di mimare e sintetizzare la capacità dell’essere umano di porsi come agente intenzionale in interazione con il proprio ambiente è data da una serie di esperimenti rispetto all’efficacia dell’implementazione del movimento biologico. Già nella descrizione dell’esperimento di Bisio et al. (2014), avevamo accennato al fatto che l’attivazione del meccanismo di risonanza motoria nell’essere umano si presentasse come strettamente condizionata dalla presenza di tratti biologici nella cinematica del comportamento. Infatti, l’esperimento ha mostrato che, se il robot performava l’azione tramite movimenti artificiali o innaturali, ad esempio non mantenendo una velocità costante del tratto agitivo accelerando bruscamente, il meccanismo di contagio motorio non veniva attivato e il partecipante ripeteva l’azione osservata tenendo in considerazione solamente l’obiettivo dell’azione, non il suo concreto svolgimento. Considerare studi che prendono in esame il ruolo dell’implementazione del movimento biologico, sviluppate e testate sul robot umanoide iCub (Metta et al. 2010), possono farci rendere conto ancora meglio del fatto che l’azione svolta tramite tempi e cinematiche di realizzazione assimilabili a quelli dell’essere umano sono un fattore cruciale per far sì che l’umano non solo adotti delle *propositional attitude* che descrivono o si rivolgono al robot in termini intenzionali, ma anche che il robot venga percepito come tale a livello corporeo.

Dunque, in questa serie di studi (Noceti et al. 2015; Sciutti et al. 2015; Vignolo et al. 2016a, 2016b, 2016c, 2017; Di Cesare et al. 2020), i ricercatori hanno preso in considerazione evidenze diffuse nelle scienze cognitive e nella psicologia dello sviluppo riguardo la capacità dei bambini in tenerissima età di discriminare il movimento biologico da quello non biologico (Simion et al. 2008; Frankenhius & Barrett 2013; Bardi et al. 2011). Ad esempio, Simion e colleghi (2008) hanno dimostrato che bambini nati da minimo 44 ore e massimo 130 ore riescono già a discriminare il movimento biologico da quello non biologico, mostrando una preferenza per il primo e per i movimenti orientati “correttamente”, ovvero quelli che mostravano figure stilizzate camminare con la testa in alto, invece che al contrario. Ovviamente, come sottolineato dagli autori, queste capacità e preferenze sono indipendenti da informazioni linguistiche o esperienze acquisite in precedenza, visto che l’esperimento è stato condotto su bambini appena nati, dimostrando come sia plausibile pensare che la capacità di riconoscere il movimento biologico sia presente già alla nascita (aprendo comunque alla possibilità che parte delle capacità possa essere sviluppata anche nell’utero e legata al principio di gravità) (Ibid.: 810).

Sembra, quindi, che la capacità di riconoscere un altro agente, come testimoniato anche in altre forme di vita animali sociali (come gatti, delfini e scimmie, cfr. Simion et al. 2008: 809), sia un elemento fondante delle abilità dimostrate più avanti nell’ontogenesi di relazionarsi agli altri. I ricercatori hanno quindi ipotizzato che dotare un robot di questa capacità avrebbe costituito un ottimo punto di partenza per rendere l’interazione tra umani e robot più fluida, in continuità con la linea di ricerca della *evolutionary robotics* (Nolfi et al. 2016; Doncieux et al. 2015; Bongard 2013): un robot che riesce a riconoscere quando si trova in presenza di un essere umano rintracciando la sua capacità di muoversi in quanto organismo biologico nel suo ambiente gli consentirebbe di intraprendere interazioni circostanziali e contestualizzate, arrivando magari fino a coordinarsi con lui (Sciutti et al. 2015).

Il problema principale che i ricercatori hanno affrontato in questo processo di implementazione consiste nella discriminazione preliminare del movimento biologico da quello non biologico. Come notato in Vignolo et al. (2016a: 1),

The examples of biological motion we face in everyday life situations is countless, and possibly characterized by very heterogeneous dynamics and trajectories. Further, even the same action may appear very different depending on the subjectivity of the moving object – each person moves in a slightly different way – or to external conditions, as light or environment, that may influence the appearance of the dynamic or constrain the user movement. The class of non-biological event is (arguably) even less constrained, because it includes all motions not produced by a living being, such as vehicles, toys, or even natural elements affected by non-biological forces (e.g., the motion of the leaves caused by the wind).

Sebbene, quindi, in termini di meccanismi di apprendimento artificiale si tratti di una classica classificazione binaria, l'eterogeneità dei dati pone un problema non banale da risolvere, che parte dalla discriminazione preliminare di un tipo di movimento dagli altri. La chiave di volta trovata dai ricercatori consiste nell'utilizzo di leggi matematiche che descrivano le caratteristiche distintive del movimento biologico, nello specifico la *Two-Thirds Power Law*. Questa legge empirica descrive una relazione costante tra la curvatura della traiettoria di un movimento e la sua velocità tangenziale (Viviani & Flash 1995). Secondo la *Two-Thirds Power Law*, la velocità del movimento lungo una traiettoria curvilinea varia in base alla curvatura: in tratti più curvi, la velocità diminuisce, mentre in tratti più rettilinei, la velocità aumenta. La formula che descrive questo comportamento è

$$v = k \cdot \kappa^{-1/3},$$

dove  $v$  è la velocità tangenziale,  $\kappa$  la curvatura, e  $k$  una costante di proporzionalità. Come dimostrato in prima istanza (Flash & Hogan 1985) nello studio in cui è stata formalizzata, questa legge descrive i tratti tipici dei movimenti biologici: i ricercatori hanno rilevato che, nel disegnare in area figure curve come un 8, la velocità del movimento non era costante, ma direttamente proporzionale al grado di curvatura della traiettoria: più la curvatura era accentuata, più il movimento rallentava, e viceversa. In sostanza, indipendentemente dalla complessità della traiettoria, esiste una relazione sistematica tra la velocità del movimento e la curvatura della traiettoria stessa, che non sembra essere specifica di un tipo di movimento in particolare quanto invece una caratteristica universale del movimento biologico. In seguito, la validità della *Two-Thirds Power Law* è stata provata in diversi esperimenti, in particolare per la percezione del movimento umano (Flach et al. 2004), per il movimento oculare (de'Sperati & Viviani 1997), per la locomozione (Hicheur et al. 2005) e per la predizione del movimento nella scrittura (Kandel et al. 2000).

A fronte di un primo studio preliminare (Noceti et al. 2015), Vignolo et al. (2016a) hanno dimostrato che la legge garantisce una classificazione robusta del movimento biologico vs movimento non-biologico, mettendola alla prova nella costruzione di un modello computazionale che «enable an artificial agent to detect the presence of humans in its surrounding to provide the appropriate pro-social behavior» (2016a: 6), confermando inoltre la possibilità di generalizzazione dei meccanismi appresi anche in scenari inediti. In particolare, il robot è stato capace di riconoscere una serie di movimenti biologici, anche molto diversi tra loro (*rolling dough* (a); *pointing* (b); *mixing a bowl* (c); *transporting an object* (d); *writing* (e)), e movimenti non biologici realizzati tramite l'impiego di oggetti di uso comune (*wheel with a random pattern* (f); *wheel with a zig-zag pattern* (g); *balloon* (h); *top toy* (i); *toy tram* (j)) (Fig. 4.5, Vignolo et al. 2016a: 3). In un altro studio (Vignolo et al. 2016b), il modello è stato implementato nel robot iCub, rendendolo capace di riconoscere il movimento dell'umano e di dirigervi la sua attenzione mentre si sta realizzando. Un terzo esperimento (Vignolo

et al. 2016c) rinforza ulteriormente l'ipotesi che la legge sia uno strumento utile nel compito di discernere movimento biologico e non biologico tramite la *Two-Thirds Power Law* e che tramite la formalizzazione computazionale in un algoritmo di apprendimento sia possibile implementare la capacità nella prospettiva di «scaling up to multiple dynamic events and to the integration of different applications» (2016c: 344).

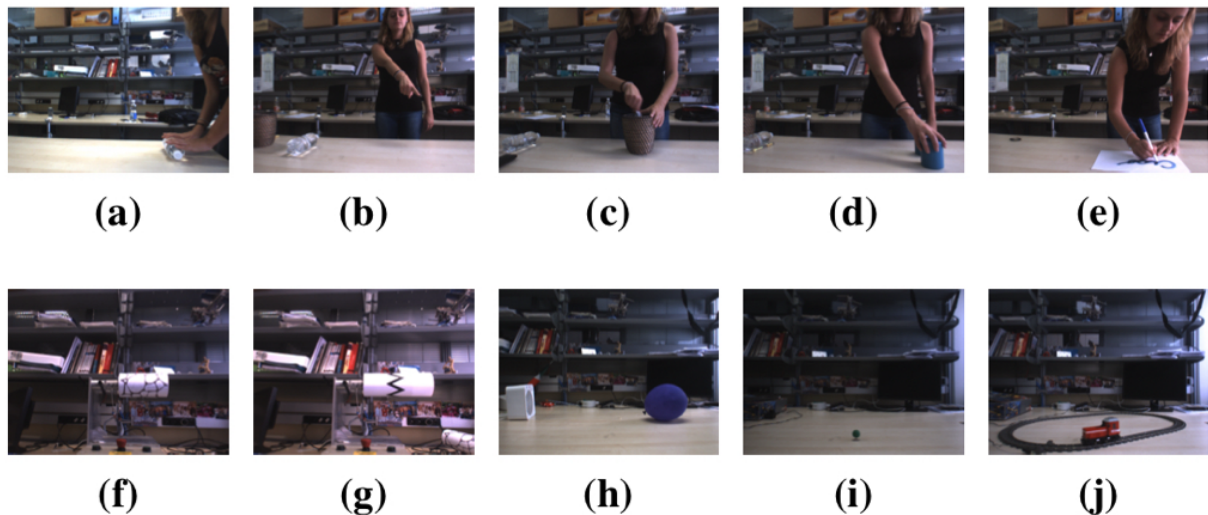


Fig. 4.5 «Biological and non-biological movements included in the training set». Vignolo et al. (2016a: 3)

Oltre all'evidenza che iCub riesca a discernere correttamente i due tipi di movimenti nei contesti quotidiani testimoniata da questi esperimenti, la validazione dei vantaggi apportati dalla implementazione delle legge riguardano anche: i) che iCub sia capace di rivolgere la sua attenzione ad altri agenti e a concentrarsi sui movimenti che compie; ii) la possibilità che iCub performi il suo movimento sulla base della stessa legge, consentendo quindi una realizzazione dell'azione che segua i principi strutturali del movimento biologico. È questo, come abbiamo visto, il tratto che più è interessante per noi, dal momento che è in questo modo che il robot riesce a mostrare un tipo di *engagement* simile a quello che un organismo biologico ha con l'ambiente all'interno del quale è inserito. Già gli esperimenti di Bisio e colleghi (2014) e di Sciutti e colleghi (2012a, 2012b, 2013, 2014) hanno fornito una sostanziale prova dell'importanza che l'azione sia performata secondo modelli biologici di movimento e intensità affinché i processi di risonanza motoria e coordinazione comportamentale fossero innescati, fornendo quindi una base solida su cui fondare l'interpretazione del robot come agente intenzionale. A questi studi si aggiunge quello di Di Cesare e colleghi (2020), i quali hanno impiegato la tecnica implicita dell'fMRI per verificare se l'area della *dorso-central insula*, attiva in contesti tipici sia nell'esecuzione sia nell'osservazione di un'azione caratterizzata da un certo tipo di *forma vitale* (ad esempio, gesto performato in maniera gentile *vs* brusca), si attivasse allo stesso modo sia di fronte all'essere umano, sia di fronte al robot umanoide che performa le azioni

con movimenti biologici. I risultati hanno mostrato che l'osservazione delle azioni performati dal robot seguendo cinematiche con un alto grado di *human-likeness* innescano l'attivazione esattamente come di fronte a un altro essere umano<sup>161</sup>. Altri studi hanno inoltre contribuito ulteriormente a rinforzare l'ipotesi che il comportamento agentivo svolto dal robot seguendo dinamiche biologiche del movimento comporti la sua considerazione come agente intenzionale, soprattutto in contesti di interazione situata (Rea et al. 2019; Kupferberg et al. 2011, 2012; Chaminade et al. 2005; Maurice et al. 2018).

Questa serie di studi dimostra una strettissima dipendenza tra la capacità del robot di performare le azioni in una modalità paragonabile al movimento biologico e la possibilità che l'essere umano lo consideri un agente intenzionale, quindi capace di interagire attivamente e di poter coordinarsi nei contesti sociali. Sciutti et al. (2015) sottolineano, inoltre, che non solamente è possibile considerare il robot come un agente intenzionale, ma anche interpretare le azioni svolte dallo stesso come compiute con specifiche intenzioni. Infatti, proprio come accade nella cognizione sociale tra umani, anche con i robot che mostrano capacità agentive all'altezza dell'interazione intersoggettiva è possibile capire se l'azione è svolta assecondando desideri o necessità, con un portato emotivo specifico, ad esempio gentilmente o in modo sgarbato (Sartori et al. 2011; Becchio et al. 2008). Nel contesto della HRI, la garanzia di questa possibilità risulta di notevole importanza, dato che permette all'essere umano non solo di comprendere le azioni che il robot sta performando all'interno della pratica interazionale che stanno avendo, ma anche di *anticipare* il modo attraverso cui il movimento verrà eseguito e il risultato dell'azione stessa.

[...] findings suggest that intentions influence action planning so that different kinematic features are selected depending on the overarching information. The observer is sensitive to this information and can use it to anticipate the unfolding of an action. Reading intention by observing movement therefore enables humans to anticipate other's actions, even when other sources of information are absent or ambiguous. (Sciutti et al. 2015: 2)

È importante, però, affinché tali dinamiche possano verificarsi, che gli esperimenti che analizzano la presenza di queste risposte motorie e interazionali nell'essere umano siano realizzati in un contesto in cui entrambi gli agenti si trovano in presenza, garantendo un incontro non più basato sulla semplice osservazione (*terza persona*) ma sull'interazione (*seconda persona*): quando l'essere umano e il robot si trovano uno di fronte all'altro, cambia anche la percezione dell'ambiente, «which is processed in

---

<sup>161</sup> Nello studio, un primo esperimento non ha mostrato attivazione della stessa area per l'osservazione del robot che performa l'azione secondo la *Two-Thirds Power Law*, ma un secondo tentativo in cui il movimento dell'essere umano era mappato tramite un tracking oculare e implementato nel robot ha mostrato il risultato opposto. Di Cesare e colleghi interpretano i dati a favore della possibilità del robot di provocare l'attivazione dell'area, dimostrando quindi che questo network non è sensibile alla differenza tra agente artificiale e agente biologico, quanto invece alla velocità e alla cinematica dell'azione svolta, a cui l'occhio umano sembra essere molto sensibile.

terms of the range of possible actions of the two partners rather than those of the single participant»; inoltre, l'interazione diretta e reciproca «is characterized by a higher flexibility, as the partners can adaptively change their actions during the interaction itself» (Sciutti et al. 2015: 3).

Un ultimo fattore da sottolineare, a cui si fa riferimento anche in questo studio, è che è importante che il robot non mostri solamente qualità biologiche nel movimento, ma anche un certo grado di somiglianza nelle fattezze. La presenza di elementi che rendano il robot anche minimamente somigliante all'umano, come la presenza di un corpo, una testa, delle braccia, delle mani, e elementi costitutivi del volto quali occhi, bocca e sopracciglia, sembrano essere fattori cruciali, ad esempio, per l'attivazione della risonanza motoria, oltre che a permettere di performare movimenti con velocità e struttura paragonabile a quelli umani: «this way, humans can easily match their own bodily configuration with that of a robot and it is also simpler for experimenters to design robot behavior approximating human motions both in end-effector and joint trajectories» (Sciutti et al. 2015: 4).

In sostanza, come documentato da questi studi, il movimento biologico implementato dai robot costituisce un elemento fondamentale che dirige l'attenzione dell'essere umano su determinate caratteristiche del suo accoppiamento con l'ambiente. L'*engagement* dimostrato dall'agente artificiale che somiglia all'umano manifesta quelle caratteristiche che fanno propendere per un'interpretazione corporea e motoria del robot come plausibile agente intenzionale. È il modo in cui il robot *mima* la capacità dell'essere umano di accoppiarsi con il proprio ambiente che innesca le dinamiche di interpretazione sociale a livello corporeo, su cui la comprensione complessiva dell'agente si struttura.

Come possiamo, quindi, definire questo rapporto tra un tipo di agentività *sintetica* e l'effetto che fa agli occhi di chi guarda? Concluderemo questo capitolo facendo riferimento a quello che abbiamo chiamato aspetto *sincretico* del robot come artefatto, al fine di determinare il tipo di intenzionalità individuata nel processo di comprensione del robot come agente intenzionale.

#### 4.4.3 *Forme di intenzionalità sincretica: fuori dai correlati ontologici*

Ora, se applicate queste considerazioni al discorso più ampio e generale dell'intenzionalità in gioco nelle interazioni sociali, possiamo ricavare una nuova visione dell'intenzionalità, non più solamente come proprietà esclusiva dell'agente avente una determinata struttura organizzativa, ma anche frutto della qualità dell'interazione che si costruisce nell'incontro stesso tra gli agenti, in cui diverse forme di agentività intenzionale sono in gioco. Come abbiamo visto in §2.2.2 e in §3.2.1, robot situati con caratteristiche interazionali come iCub (Metta et al. 2010) hanno la capacità di instaurare dinamiche di dialogo e dispiegare una serie di azioni nel mondo che, anche se dalla struttura elementare e dalla complessità ridotta, costruiscono un dominio relazionale che in qualche modo può

essere definito sociale. In qualche modo, quindi, robot come questo sono dotati di capacità sensomotorie che allestiscono relazioni tra l'*embodiment* (organismoide, cfr. §3.4.2) del robot e l'ambiente che anche noi condividiamo: iCub ha la capacità di seguire lo sguardo del suo interlocutore, di esibire espressioni facciali anche se minime, di *muoversi* in accordo alle dinamiche e alla cinetica dell'essere umano. Di comporre, in sostanza, un comportamento.

Since all the movements of the organisms are always conditioned by external influences, one can, if one wishes, readily treat behavior as an effect of the milieu. But in the same way, since all the stimulations which the organisms receives have in turn been possible only by its preceding movements which have culminated in exposing the receptor organ to external influences, one could also say that *behavior is the first cause of all the stimulations*. (Merleau-Ponty citato in Varela 1992b: 104)

Varela usa questo passo da *La struttura del comportamento* di Merleau-Ponty (1942) per spiegare come «perception is not simply embedded within and constrained by the surrounding world; it also contributes to the *enactment* of this surrounding world» (Varela 1992b: 104). In questo senso, l'affermazione di Varela parte dall'idea espressa dal fenomenologo secondo cui il comportamento non è una risposta a stimoli esterni raccolti tramite percezione, arrivando ad affermare uno dei principi fondativi della teoria dell'enazione: l'*engagement*. In altre parole, gli organismi partecipano attivamente tramite le loro stesse azioni alla creazione dell'ambiente, il quale si correda di *affordance* che emergono a causa dell'accoppiamento strutturale con l'*embodiment* dell'organismo. Il comportamento esibito da un agente, dunque, non è soltanto l'effetto di risposta al contesto e al mondo esterno, ma è la causa che struttura quel mondo, innescando un *loop* di influenze dinamiche in cui abito un ambiente che mi modifica, e modifico l'ambiente in cui abito. Noi, d'altra parte, possiamo leggere in questo passo una conferma delle ipotesi che abbiamo cercato di provare in questo capitolo, andando per punti:

- In primo luogo, abbiamo voluto mettere in discussione l'approccio maggioritario dell'attribuzione di intenzionalità, dal momento che abbiamo riscontrato inconsistenze tra gli approcci ibridati e una interpretazione in alcuni dei contesti sperimentali basata su cortocircuiti teorici dovuti alla componente di ToM inclusa.
- Abbiamo, quindi, proposto una nozione di intenzionalità che non presentasse conflitti e incongruenze interne, in cui alcuni degli approcci classici (neo-comportamentista e neo-pragmatista) potessero convivere e nutrire il concetto di intenzionalità enattivista. Il tipo di intenzionalità proposta, inoltre, ci ha consentito di illuminare un ulteriore aspetto, corporeo e sensorimotorio, da indagare al fine di comprendere se il robot è considerato un agente sociale in interazione. Gli studi che abbiamo chiamato in causa dimostrano una robusta corrispondenza tra risposte comportamentali, motorie e neurocognitive e considerazione del robot come agente sociale in contesti interazionali.

- Tuttavia, la concezione dell'intenzionalità originale nell'enattivismo, dato lo stretto legame con il *sense-making*, ci ha condotto a specificare ulteriormente il processo che conduce l'essere umano a trattare il robot come agente intenzionale, anche se questo non può essere considerato produrre le stesse capacità di costruzione di significato rispetto a un organismo vivente. Siamo ricorsi, dunque, agli studi sull'implementazione del movimento biologico per mostrare che l'attenzione in questa indagine non debba concentrarsi esclusivamente sulle qualità della struttura organizzazionale del sistema che producono come effetto la possibilità di un *sense-making* e un'intenzionalità genuina, ma sul rapporto che l'organismo intrattiene con il proprio ambiente tramite espressioni agentive di intenzionalità incarnata, ovvero sull'*engagement*.

Siamo arrivati, ora, a dover descrivere il tipo di intenzionalità in gioco in questo tipo di interazioni, realizzata dal robot e rilevata, individuata, riconosciuta (*detected*) dall'essere umano, il quale reagisce in modo sociale e intercorporeo alle azioni del robot stesso. Crediamo che il modo migliore per farlo sia ricorrendo al concetto di *sincretismo* esposto nel capitolo precedente (§3.5).

Quando abbiamo descritto la struttura *mimetica* e *sintetica* dell'artefatto robotico (§3.4), ci siamo soffermati sul fatto che queste due caratteristiche non possano essere considerate presentarsi in isolamento, né l'una dall'altra, né rispetto al contesto in cui l'interazione avviene. Questo ci ha portato a considerare, quindi, il robot innanzitutto come artefatto *sincretico*, ovvero come agente che trova il suo potenziale sociale non nella realizzazione di un meccanismo computazionale efficace perché condiviso con gli esseri umani, ma perché si colloca come interattante nell'anello interazionale, in cui l'intreccio di capacità mimetiche e sintetiche enattiva gradienti di socialità e in cui l'interazione gioca un ruolo cruciale per definire i parametri dell'aggiustamento reciproco e coordinato. Il modo attraverso cui questo concetto può illuminare la struttura dell'intenzionalità che abbiamo rintracciato è almeno duplice:

- i. Da una parte, ci fa comprendere come la componente estetica, dai più considerata come il fattore scatenante dell'attribuzione di intenzionalità ai robot (Epley et al. 2007; Spatola & Wykowska 2021; Spatola et al. 2021a; Marchesi et al. 2021a), non sia che un tassello in quadro più grande nel fenomeno dell'*intentionality detection*, che comprende anche capacità sintetiche di realizzazione del movimento biologico.
- ii. Dall'altra, il concetto di *sincretismo* ci consente di considerare come sia prima di tutto l'interazione che si costruisce a guidare l'interpretazione del robot come agente sociale portatore di un tipo di intenzionalità diversa da quella originaria intrinseca ai processi di *sense-making* dell'organismo.



Quello che vogliamo proporre, allora, è di considerare il tipo di intenzionalità realizzata dal robot in interazione sociale, che gli permette di essere considerato un agente che realizza le proprie azioni come rivolte a uno scopo, come una forma di *intenzionalità sincretica*. L'intenzionalità sincretica è ciò che sussume gli aspetti *mimetici* e *sintetici* del robot come agente intenzionale: da una parte, il fatto che realizzi una corporeità identificabile come simile a quella umana comporta la sua capacità di prendere il posto di un agente (Airenti 2018; Damiano & Dumouchel 2018), assumendo quindi i diritti e i doveri di un agente intenzionale; dall'altra, la componente sintetica che permette al robot di realizzare una capacità motoria simile all'umano (*embodiment organismoide*, Ziemke 2003a), con movimenti che riproducono la stessa *ratio* tra velocità e costanza della traiettoria del movimento (Bisio et al. 2014; Noceti et al. 2015; Sciutti et al. 2015; Vignolo et al. 2016a, 2016b, 2016c, 2017), garantisce un trattamento particolare tra gli arti artefatti, che lo fa apparire, agli occhi e al corpo di chi guarda, un agente intenzionale pronto a interagire (Sciutti et al. 2012a, 2013).

In questo quadro, il comportamento realizzato dal robot va visto non solo nella sua componente corporea, ma nel suo legame con l'ambiente che lo accoglie: «[w]e must see the organism and environment as bound together in reciprocal specification and selection» (Varela 1992b: 104). È solo così che l'intenzionalità del robot può emergere agli occhi di chi guarda come un tratto del comportamento concretizzato in accoppiamenti agentivi dell'agente che lo realizza. L'idea trova perfetta sovrapposizione con l'approccio dinamico nelle scienze cognitive, che concepisce il comportamento esattamente come una proprietà dell'intero sistema accoppiato composto da cervello, corpo e mondo: «[b]ehavior is a property of an entire BBE [Brain-Body-Environment] system and cannot in general be properly attributed solely to any one component of that system» (Beer 2023: 7). Lo stesso vale, allora, per l'intenzionalità, che non può essere localizzata in un tratto specifico di questo accoppiamento, ma va sempre vista come una qualità emergente dello stesso. Possiamo rileggere così le intuizioni di Dennett quando parlava, commentando le idee di Elizabeth Anscombe, dell'impossibilità di localizzare l'intenzionalità in un tratto specifico del comportamento dell'agente (Dennett 1987: 61, nota 7). Anche per Dennett, quindi, l'intenzionalità non si dà come capacità identificabile come marchio del mentale interno all'agente, ma come qualità data dalle possibilità di generazione di un comportamento: «intentionality doesn't come from on high; it *percolates up from below*» (Dennett 1995: 205, corsivi miei).

Per riprendere ancora un'idea proveniente dalle teorie dei sistemi dinamici, l'effetto dato da questo accoppiamento tra corpo, cervello e mondo – ciò che *percola* in su dal basso – è, solitamente, ciò che vediamo compiersi quando osserviamo il dispiegarsi del comportamento. L'accoppiamento dinamico che genera le capacità cognitive non si limita a operare internamente, ma si esprime attraverso

l'interazione attiva con l'ambiente: in altre parole, l'efficacia cognitiva di un organismo (o di un artefatto) non è rilevante finché non viene tradotta in azione.

Aside from emulating a rock, the only way for an organism to persist is to actively engage its environment, struggling for the resources it requires to survive and opposing the many forces that would disintegrate it. The biological purpose of cognition is to contribute to the *behavior* that is absolutely essential to an organism's success (Lyon 2006). To a cognitive ear, this emphasis on behavior will sound very much like behaviorism, but it is not (Barrett 2012). It is merely a recognition of the fact that it is only the manifestation of internal processes in overt action that has actual consequences in the world and can, therefore, be subjected to selection. Internally, rocks may be the most intelligent beings in the universe, but the rest of the universe will never know this if they just sit silently for all eternity. (Beer 2023: 7)

L'intenzionalità sincretica, allora, è proprio ciò che risulta dall'osservazione non di un processo genuino e dinamico di messa in relazione dell'organismo con il proprio ambiente, ma solamente degli effetti che questo produce. L'essere umano *percepisce* il robot come possibile unità dialogante, come partner nell'interazione a due, come qualcosa con cui può costruire un piano della comunicazione a differenza di quanto avviene con bambole, dipinti e altre simili riproduzioni dell'essere umano. È mimando il diagramma formato dal comportamento di un agente intenzionale genuino nel proprio ambiente che il robot allestisce la possibilità di essere considerato agente intenzionale; è sintetizzando le proprietà cinematiche del movimento biologico che può rinforzare il proprio legame con l'interazione sociale. Ma, come abbiamo visto in precedenza, queste due caratteristiche sono discernibili solamente a posteriori, dal momento che prima di tutto il robot le presenta insieme: nel sincretismo del porsi come interattante allo sguardo dell'osservatore e realizzare un'armonia degli effetti.

Possiamo quindi riassumere le caratteristiche dell'*intenzionalità sincretica*, così da renderle operative e considerarne i vantaggi:

- L'intenzionalità sincretica è una proprietà dell'*engagement* del robot con il suo ambiente, che si genera per causa dell'interazione dinamica e situata che si configura come sociale. Il robot sfrutta i processi dinamici che l'essere umano mette in atto nell'interazione sociale sia sintetizzando le sue capacità di muoversi nel mondo, sia quando si trova di fronte a lui: si affida alle inerenti capacità sociali dell'agente umano e si inserisce nel circolo dell'esercizio delle abilità di dare senso al mondo che il robot stesso non possiede, ma che abilmente abita. Potremmo dire che il robot sfrutta gli *abiti* (§5.3) incarnati che l'essere umano riversa nell'interazione nel momento stesso in cui si costituisce<sup>162</sup>.

---

<sup>162</sup> Possiamo anticipare quello che diremo in §5.3 attraverso questo passaggio di Gallagher (2017: 80): «My own intentions are operable, and quite often only emerge, within my perception of the other's intentions. I see the other's actions as an affordance for my own possible responsive action [...]; I see the other's action as inter-actionable or as calling forth a response on my part».

- L'effetto prodotto negli occhi di chi guarda è, letteralmente, lo stesso di un essere umano: l'intenzionalità sincretica è caratterizzata da una risposta corporea indistinguibile nell'essere umano da quella enattivata in contesti di interazione sociale intersoggettiva. Il tema dell'intenzionalità attribuita, o non-originale, quindi, prende una piega diversa da quella che abbiamo rintracciato negli altri approcci: qui, questo tipo di intenzionalità derivata, poiché non correlata ad alcuna capacità di *sense-making* genuina, non è qualcosa che proviene esclusivamente dall'agente che osserva il robot in maniera distaccata – in terza persona – ma si configura come proprietà dell'interazione, la quale distribuisce salienze percettive e agentive sulla base delle possibilità di co-costruzione di senso messe in gioco. Nemmeno l'intenzionalità dimostrata dall'essere umano, d'altronde, è esclusiva proprietà del suo corpo o delle narrazioni che usa per interpretare il mondo in maniera linguisticamente strutturata; anche nell'essere umano, l'intenzionalità è la proprietà esibita dall'*engagement* dispiegato tra l'organismo e il suo ambiente.

Quello che teniamo a sottolineare è che il grado di sincretismo raggiunto dal robot, che ha come uno dei suoi effetti un'intenzionalità dello stesso tipo, si costruisce a partire dalla capacità dell'essere umano in interazione di costruire piani di senso condivisi con l'interlocutore e dalla struttura intrinsecamente partecipativa della cognizione sociale umana. Le operazioni che il concetto di *intenzionalità sincretica* ci permette consistono in due azioni principali:

- Ci permette di capire come sia possibile avere un'interazione sociale con un agente che, se considerato in isolamento, nella struttura che permette di dimostrare qualche forma di agentività, non possiede alcuna capacità autonoma di auto-determinarsi, di dare senso al mondo e di ingaggiare dialoghi interazionali sufficientemente complessi. Come dimostrato nell'analisi di Di Paolo (2003), la composizione strutturale e organizzazionale del robot non ha la possibilità di realizzare forme genuine di *sense-making*, di mostrare intenzionalità non-derivata o di coordinarsi in interazione in modo consapevole; e tuttavia, stando agli esperimenti e alle evidenze della HRI, l'essere umano lo percepisce a livello corporeo come un partner interazionale. L'intenzionalità sincretica si delinea quindi come quell'effetto che permette di considerare il robot non come un agente intenzionale genuino, ma gli riconosce una capacità che si manifesta agli occhi di chi interagisce con lui.
- Considerare l'intenzionalità sincretica nella HRI ci consente di concentrare l'attenzione sia sul contributo che i due agenti individualmente apportano all'interazione, sia vedere come sia l'interazione stessa, spesso, a determinare lo statuto interazionale degli agenti coinvolti. In questo senso, i processi percettivi e sensorimotori che permettono la rilevazione di forme di

intenzionalità sincretica debbono essere riletti alla luce del contributo dell'interazione alla strutturazione del piano percettivo. Come sottolineato da De Jaegher, «we are no longer tied to thinking of intentions as individually formed and then brought into a social situation to be interpreted, simulated, or directly perceived» (De Jaegher 2009: 540): «More intricate possibilities are also opened up: we can now think about intentions can be transformed – even generated – in social interaction» (2009: 539). Alla fine del percorso sul ruolo dell'intenzionalità nella HRI possiamo dire non solo che questo è possibile, ma che l'interazione stessa permette a un agente, che intrinsecamente non sarebbe in grado di sviluppare alcuna forma di intenzionalità, di acquisire intenzionalità proprio attraverso lo scambio sociale attivo e le dinamiche di mutuo riconoscimento che si verificano a livello corporeo. In sostanza, l'intenzionalità sincretica ci consente di andare oltre l'intenzionalità individuale, non solamente prendendo in considerazione il modo in cui gli agenti coinvolti modulano reciprocamente le proprie intenzionalità genuine – come, ad esempio, può avvenire nei fenomeni di *shared intentionality* (Tomasello & Carpenter 2007) o di *we-intentionality* (Zahavi 2021) – ma rendendo possibile comprendere come anche un agente che intenzionalità genuina non possiede possa arrivare a entrare a far parte di dinamiche sociali e affettive e innescare processi sincretici di mutua interpretazione.

Infine, l'intenzionalità sincretica ci permette di dare una nuova lettura alla frase con cui siamo partiti, in cui Simondon (1958: 11) descrive il robot come qualcosa che non esiste individualmente, *qu'il n'est pas une machine*, ma solamente come prodotto dell'immaginazione e della fabbricazione di fantasia. Ci troviamo, ora, a confrontare questa visione con la realtà delle interazioni moderne tra umani e robot, dove il concetto di artefatto sociale si è evoluto ben oltre l'antico costrutto mitologico. Tuttavia, Simondon pare aver compreso già prima degli sviluppi tecnologici della robotica sociale che ciò che consente al robot di ricoprire quel ruolo speciale nell'insieme degli artefatti antropomorfi è una costitutiva relazione con l'umano, in cui acquista la possibilità di sembrare più di quello che strutturalmente è. Il robot non trova la sua identità in un'entità separata dall'umano, ma come interattante che possiede capacità mimetiche e sintetiche tali da enattivare forme di socialità attraverso l'interazione dinamica con l'ambiente e l'altro agente. Possiamo spostare il focus, quindi, da una concezione ontologica individualistica a una più relazionale e distribuita, in cui l'intenzionalità emerge dalle contingenze dell'interazione situata. L'intenzionalità sincretica, lungi dall'essere semplicemente una proiezione antropomorfica, è il risultato di un processo co-enattivo che si sviluppa nell'incontro tra umano e robot. Ci troviamo ancora d'accordo con Simondon, quindi, nel dire che, come entità individuale, il robot non esiste: attraverso le sue azioni e la sua capacità di rispondere al

flusso interattivo, non esiste più come singolo oggetto tecnico, ma si colloca in un dominio relazionale che genera nuove forme di agentività e di intenzionalità che si concatenano alle logiche del *sense-making*.

La domanda che a questo punto si pone è: come è possibile, allora, che, nonostante la capacità dei robot di enattivare risposte sensorimotorie così simili a quelle riscontrate nell'interazione sociale con altri umani, l'interazione con loro si presenta, ancora oggi, così problematica? Numerose, infatti, sono le testimonianze di rottura della HRI, che si manifestano nell'essere umano come delle sensazioni spiacevoli, di disagio, le quali a volte possono anche condurre a volersi allontanare dal robot stesso e interrompere qualsiasi forma di collaborazione o vicinanza con l'artefatto. Per cercare di capire come sia possibile che fenomeni come questo avvengano e quali siano i processi che conducono alla disgregazione del regime interazionale in cui la natura sincretica del robot si frantuma, nel prossimo capitolo prenderemo in considerazione il fenomeno dell'*Uncanny Valley Effect* (Mori 1970) e, analizzandone la struttura e considerando gli studi fino ad ora condotti sul tema, forniremo un modello di interazione sociale basato sui principi enattivisti che potrebbe aprire nuove strade di collaborazione e mutua comprensibilità con gli artefatti robotici.

**Parte III –**  
**Balla coi robot. L'alterità della macchina**

## Capitolo V – Il gioco dell'interazione.

### Rotture, abiti e aspettative nella HRI

It might be urged that when playing the 'imitation game' the best strategy for the machine may possibly be something other than imitation of the behavior of a man.

This may be, but I think it is unlikely that there is any great effect of this kind.

(Turing 1950: 435)

#### 5.1 Tra imitazione e interazione: l'enigma dei robot sociali

Nel primo capitolo I (§1.2.4), abbiamo parlato del *gioco dell'imitazione* come strumento teorico sviluppato da Alan Turing (1950) allo scopo di rispondere alla domanda «Can machines think?» (1950: 433) e abbiamo visto come questo abbia aperto a una serie di riflessioni filosofiche che, ancora oggi, sono al centro del dibattito sull'intelligenza artificiale e nelle scienze cognitive. Abbiamo visto, inoltre, come il test di Turing abbia nutrito l'ambizione dell'umano di costruire repliche di sé stesso, e come questo abbia contribuito a far evolvere ancora una volta il concetto di macchina che avrebbe costituito le fondamenta dell'impresa dell'intelligenza artificiale e della robotica. Ora, alla luce degli sviluppi che ci hanno condotto fino a qui e con le adeguate variazioni metodologiche ed epistemologiche, il gioco dell'imitazione può fornire una base per farci vedere meglio la struttura costitutiva dell'interazione tra umani e robot.

Il primo aspetto che prenderemo in considerazione del gioco dell'imitazione è la componente mimetica-comportamentale considerata da Turing alla base del test. Nell'articolo del 1950, Turing apre a un'interpretazione delle *performance* della macchina computazionale non necessariamente legata alla composizione interna dei meccanismi che generano il suo comportamento, ma alla sostanziale efficacia che il comportamento stesso assume all'interno dello scambio con l'interrogatore umano, che nel caso del test di Turing – ricordiamo – si presenta come un'interazione dialogica di tipo verbale<sup>163</sup>. Ciò che conta, dice Turing, è che il comportamento sia realizzato in maniera tale da

---

<sup>163</sup> Facciamo riferimento alla formulazione originale del test, descritta in Turing (1950). Nella prima versione (1950: 433-434), il test prevede che ci siano tre attori: un uomo (A), una donna (B) e un interrogatore (C). C è posizionato davanti a una *telescrivente* in una stanza, mentre A e B sono in un'altra. Lo scopo del gioco varia: per l'interrogatore consiste nel determinare quale dei due attori nell'altra stanza è un uomo e quale una donna, ponendo domande attraverso il supporto di scrittura ad A e B; per A, lo scopo è far sbagliare l'interrogatore C; per B, lo scopo è aiutare C a fare la scelta giusta. L'ipotesi da verificare nel test è: «What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?» (Turing 1950: 434). Nel corso dei decenni, il test dell'imitazione ha subito molte modifiche (French 2000;

essere considerato interscambiabile con quello di un essere umano, almeno per un lasso di tempo minimo.

May not the machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection. (Turing 1950: 435)

Abbiamo visto come questa posizione abbia costituito la base del principio di realizzabilità multipla al cuore del funzionalismo cognitivista (Putnam 1960) e come abbia nutrito e rafforzato l'idea secondo cui, se la macchina fosse stata capace di realizzare gli stessi effetti del pensiero umano nello svolgimento di un compito cognitivo, allora si sarebbe avuta prova della struttura computazionale della mente (Fodor 1975). A fronte delle argomentazioni che abbiamo sviluppato, e rileggendo la proposta di Turing riguardo la possibilità della macchina di emulare alcune funzioni del pensiero umano, possiamo riconsiderare i principi del gioco dell'imitazione evitando di trarne delle conseguenze epistemologiche sulla strutturazione interna delle funzioni cognitive, ma focalizzando l'attenzione sul ruolo giocato dall'interazione nella realizzazione del test. Il percorso che abbiamo compiuto, infatti, ci consente di rileggere il test di Turing non più (o non solamente) come la plausibile prova dell'omologazione dei meccanismi computazionali a quelli del pensiero umano, ma come un test per verificare la capacità della macchina di co-creare significati attraverso l'interazione con l'interlocutore umano.

In questo capitolo, dunque, vogliamo sviluppare una variante del gioco dell'imitazione proposto da Turing evidenziando il ruolo della componente dialogica e partecipata dello scambio che avviene con la macchina: non chiederci più “Can machines think?”, ma “Can robots interact?”. Il gioco, quindi, da essere basato sui principi imitativi che possono essere incorporati dalla macchina, diventa quello di reggere il peso della scambievolezza e della reciprocità: *il gioco dell'interazione*, un gioco che noi esseri umani giochiamo fin dalla nascita, in cui lo scopo è coordinarsi agli altri all'interno di un panorama di significati socialmente distribuiti, che si struttura a partire dalle dinamiche interazionali sensorimotorie basate soprattutto su comprensioni incarnate e tenute pragmatiche dei significati condivisi. Lo scopo del gioco non è allora più quello di decretare la possibilità che i meccanismi implementati nel robot siano prova dell'ontologia organizzazionale e cognitiva comune a esseri umani e a macchine, ma che si possa sondare la possibilità del robot di entrare nella co-costruzione dinamica del senso attraverso la prova dell'interazione.

---

Searle 2009; Hernández-Orallo 2020), sviluppando anche versioni di interazione non verbale (ad es.: Ciardo et al. 2022; Pfeiffer et al. 2011)



Per capire meglio in cosa consiste questa nuova variante del gioco che proponiamo, possiamo innanzitutto guardare a due delle maggiori differenze con il gioco dell'imitazione, entrambe conseguenze del percorso che ci ha condotto fino a qui:

- La prima differenza è riscontrabile nell'obiettivo del gioco: da una parte, lo scopo del gioco dell'imitazione è testare se una macchina possa imitare il comportamento linguistico dell'essere umano, in modo così accurato da risultare indistinguibile a una prova interlocutoria "cieca"; nel gioco dell'interazione, l'obiettivo è quello di verificare se il robot è in grado di soddisfare le esigenze interindividuali date da un'interazione continua e contestuale con un essere umano, non limitandosi dunque all'imitazione ma includendo processi *mimetici* che riguardano la necessità di coordinazione, di adattamento, e la co-costruzione dei significati comuni che si sviluppano nel tessuto dinamico della mutualità.
- In secondo luogo, i giochi differiscono nelle modalità di realizzazione: se il gioco dell'imitazione si svolge in un contesto statico e disincarnato, in cui gli interlocutori non comunicano se non attraverso lo schermo della telescrivente e non si vedono mentre il gioco si dispiega, il gioco dell'interazione, invece, prevede necessariamente che l'essere umano possa almeno vedere l'agente muoversi nel suo ambiente, anche se non è condiviso con il suo: se non necessariamente l'effettiva presenza di entrambi gli agenti nello stesso luogo, ciò che si richiede è perlomeno che sia possibile osservare il robot nello svolgimento di un'azione.

Prima di tutto, dobbiamo dunque chiederci se la macchina che stiamo considerando – il robot sociale umanoide – possieda le carte in regola per candidarsi come partecipante al nuovo gioco che stiamo delineando. Durante il nostro percorso, abbiamo visto come quelle caratteristiche che compongono l'intenzionalità sincretica del robot permettano a quest'ultimo di presentarsi nell'interazione con l'umano come un agente intenzionale e di enattivare risposte correlate nell'essere umano. Sviluppata grazie alle capacità agentive dimostrate dal robot nel proprio ambiente, infatti, l'intenzionalità sincretica è ciò che permette al robot di essere considerato un agente *potenzialmente* capace di interazioni intersoggettive, ovvero di assurgere al ruolo di interattante nello spazio fisico dell'incontro con l'altro. Dunque, entrambi gli aspetti raccolti nella presentazione *sincretica*, ovvero la caratterizzazione *mimetica* che ne istituisce la somiglianza all'umano e quella *sintetica* che ne orienta il comportamento, collocano immediatamente il robot umanoide in un campo di gioco in cui ciò che conta è la capacità di essere partner interazionale per qualcuno.

In questo senso, il gioco dell'interazione ha in comune almeno due aspetti fondamentali con il gioco dell'imitazione:

- In primo luogo, entrambi i giochi puntano a valutare l'adeguatezza della macchina in contesti di scambio interazionale. Sebbene attraverso strutture diverse, lo scopo dei giochi, infatti, è quello di misurare alcune capacità degli artefatti artificiali sulla base della performance svolta in interazione con gli umani: che sia tramite il dialogo o tramite l'interazione *embodied*, il banco di prova che stabilisce se la macchina è all'altezza dell'umano consiste in un'interazione viva e nell'esercizio delle capacità comportamentali che costruiscono l'interazione sociale.
- In secondo luogo, per entrambi si configura un'attenzione particolare ai momenti di *rottura*. Sia nel gioco dell'imitazione, sia in quello dell'interazione, il momento in cui l'essere umano si rende conto della instabilità dello scambio, della distanza tra ciò che dovrebbe essere e quello che effettivamente si realizza, segna il punto in cui il gioco si rompe e il test fallisce: ovvero, il momento in cui è più difficile recuperare, che spesso ha come esito l'eclissi delle possibilità di confronto, quello in cui i legami reciproci diventano univoci e si sfilacciano a mostrare il flusso grezzo della macchinosità dell'interazione, che fino a poco prima scorreva fluido nel terreno delle aspettative non ancora messe alla prova.

Ora, proprio la *rottura* del processo di scambio dialogico tra umani e macchine è quello che meglio può illuminarci un cammino che risulta ancora intentato nella comprensione dell'interazione tra umani e robot. Già nel capitolo precedente abbiamo sottolineato la centralità dei momenti di interruzione dell'interazione: seguendo l'impostazione enattivista all'intenzionalità e alla sua *detection*, abbiamo visto come le rotture interazionali fossero il momento in cui, a forme di comprensione motoria e incarnata che contraddistinguono l'*intentional attitude*, vengono in soccorso *propositional attitude*, pezzi del panorama narrativo e culturale organizzati linguisticamente atti a rinforzare il processo di intendimento con l'altro in situazioni di incertezza interpretativa: quando devo aiutare un mio amico a capire perché la sua fidanzata lo ha lasciato, quando devo interpretare una situazione in cui non sono coinvolta o che presenta tratti di complessità, o quando le mie aspettative corporee non sono soddisfatte in interazione. Un'analisi, quindi, del momento in cui l'interazione inizia a incrinarsi, e delle maniere in cui si tenta di recuperarla in qualche modo, può aiutarci a delimitare il fenomeno della comprensione del robot in interazione. Cercheremo di rintracciare le motivazioni delle rotture e di dare forma al processo di recupero della comprensione che ne segue attraverso lo sguardo enattivista alla HRI che abbiamo costruito nei capitoli precedenti, testando questa volta il modello su ciò che accade nell'interazione nel suo complesso. Come si struttura l'interazione tra umani e robot? Quali sono i processi che possono aiutarci a capire in che modo descriverla al fine di ottimizzarla?

Vogliamo partire da questo ultimo punto, allora, relativo al ruolo della rottura dello scambio nel gioco dell'imitazione, per definire la sfida che ci spetta in quest'ultimo capitolo, analizzando il valore del *breakdown* in ottica enattivista affinché sia possibile rendere conto di come comprendere alcuni fenomeni tipici della HRI e, da questo, ricavarne delle opportunità di comprensione laddove, spesso, viene sancita la fine dei giochi.

#### 4.4.4 Il valore della rottura: per un'analisi dei *breakdown*

Abbiamo già incontrato, nel corso della presente trattazione, un approccio alle scienze cognitive e alla costruzione di artefatti artificiali che ponesse l'enfasi sui momenti di rottura della pragmatica dell'interazione. Nel capitolo II (§2.1.1.2), infatti, abbiamo visto come i *breakdown* dell'interazione siano un elemento cardine per un'analisi dell'esperienza di fronte alle macchine intelligenti all'interno dell'approccio presentato da Terry Winograd e Fernando Flores (1986), il primo nella corrente *mainstream* a impostare un legame diretto tra la teoria dell'autopoiesi di Maturana e Varela (1980) e le tecniche di intelligenza artificiale. Nella definizione data da Winograd e Flores,

By [breakdown] we mean the interrupted moment of our habitual, standard, comfortable 'being-in-the world'. Breakdowns serve as an extremely important cognitive function, revealing to us the nature of our practices and equipment, making them 'present-to-hand' to us, perhaps for the first time. In this sense they function in a positive rather than a negative way. New design can be created and implemented only in the space that emerges in the recurrent structure of breakdown. A design constitutes an interpretation of breakdown and committed attempt to anticipate future breakdowns. (Winograd & Flores 1986: 77-78)

A partire dalla commistione istituita tra la *readiness-to-hand* di Heidegger e la teoria dell'autopoiesi per come elaborata da Maturana<sup>164</sup>, Winograd e Flores avevano individuato nei momenti in cui l'interazione con la macchina non funziona, ovvero non dà i risultati sperati dal costruttore o programmatore, l'occasione di testare la discrepanza tra intento d'uso e pratica di utilizzo.

When we look at computer technology this basic point guides our understanding in several ways. First, it is critical in our anticipation of the kinds of computer tools that will be useful. In a tradition that emphasizes thought as an independent activity, we will tend to design systems to work within that domain. In fact much of the current advertising rhetoric about computers stresses the role they will play in 'applying knowledge' and 'making decisions'. If, on the other hand, we take action as primary, we will ask how computers can play a role in the kinds of actions that make up our lives – particularly the communicative

---

<sup>164</sup> «For Heidegger, 'things' emerge in breakdown, when unreadiness-to-hand unconceals them as a matter of concern. Maturana sees the presence of objects and properties as relevant only in a domain of distinctions made by an observer. In the domain of biological mechanism, they do not exist. Both authors recognize that we are situated in a world that is not of our own making. Their central insight is that this world, constituted as a world of objects and properties, arises only in the concerned activities of the person». (Winograd e Flores 1986: 73)

acts that create requests and commitments and that serve to link us to others. (Winograd e Flores 1986: 71-72)

Il criterio, quindi, non risiede più nelle omologazioni tra processi computazionali ritenuti essere alla base della struttura cognitiva tanto dell'umano quanto del computer, ma nella (dia-)logica dell'azione, che si dispiega nella pratica d'uso attiva. In questo contesto, i momenti di rottura non vanno considerati come qualcosa di esclusivamente negativo, che impedisce il corretto svolgimento del compito cognitivo e il raggiungimento dell'obiettivo, ma come un'occasione che deve essere osservata, considerata nel suo complesso e analizzata proprio perché può svelare aspetti da migliorare: «A breakdown reveals the nexus of relations necessary for us to accomplish our task» (Winograd e Flores 1986: 165).

A seguito di questa formulazione fornita da Winograd e Flores, e soprattutto in corrispondenza del ribaltamento di prospettive che proponeva la riconsiderazione della cognizione da esclusivamente razionale e linguistica a incarnata e rivolta all'azione, l'importanza dei *breakdown* nell'interazione è stata riconosciuta da Varela (1992b) come il punto d'accesso alla comprensione di momenti particolari nel processo di cognizione umana. Considerare il modo in cui il filosofo cileno ha declinato questa impostazione in riferimento ai *breakdown* può farci meglio comprendere quanto l'analisi di questi momenti possa rivelare aspetti importanti dell'interazione, tanto dell'organismo con il suo ambiente, quanto dell'organismo con altri agenti immersi nel suo stesso mondo.

Nell'idea di Varela, il *breakdown* costituisce un passaggio, una zona locale di condensazione di senso che intercorre tra due distinte e definite situazioni, ognuna con i suoi principi regolativi e i suoi sviluppi identificabili, intervallate da un momento di esitazione. Il modo in cui Varela descrive l'emersione di una rottura definisce chiaramente i termini in gioco:

Picture yourself walking down the street, perhaps going to meet somebody. It is the end of the day and there is nothing very special in your mind. You are in a relaxed mood, in what we may call the readiness of the walker who is simply strolling. You put hand into your pocket and suddenly you don't find your wallet where it usually is. *Breakdown*: you stop, your mind setting is unclear, your emotional tonality shifts. Before you know it, a new world emerges: you see clearly that you left your wallet in the store where you just bought cigarettes. Your mood shifts now to one of concern for losing documents and money, your readiness-to-action is now to quickly go back to the store. There is little attention to the surrounding trees and passersby; all the attention is directed to avoiding further delays. (Varela 1992b: 101, corsivi miei).

Ciò che si delinea è allora una struttura processuale aperta all'incertezza, che condensa le contestuali possibilità di azione aperte dalla rottura di un *frame* interazionale nel momento esatto della presa d'atto che c'è *qualcosa che non va*.

L'esempio riportato da Varela è esaustivo e non si fa fatica a tradurlo in termini di incontri con altri agenti all'interno del nostro mondo sociale. Si pensi, ad esempio, alla situazione in cui siamo a una conferenza e stiamo per presentarci a qualcuno per la prima volta. Il contesto è formale e questo ci fa

selezionare una serie di pratiche utili che instradano il nostro comportamento in questa determinata situazione, in cui dobbiamo presentarci a qualcuno: d'istinto, quindi, allungheremo la mano verso l'altra persona e attenderemo che la stringa, sancendo così l'inizio dell'interazione. Per qualche motivo, però, questo gesto non va come previsto: forse il nostro partner interazionale non ha una presa salda o non sta vedendo la nostra mano tesa verso di lui, oppure ritrae la sua improvvisamente prima di afferrare la nostra. Fatto sta che il gesto non si compie e qualcosa si rompe. Almeno una volta nella vita chiunque ha vissuto la situazione di imbarazzo provocata dalla incomprensione di un gesto eseguito, magari a causa della nostra esecuzione o delle molteplici possibilità di significazione che il gesto stesso ha in interazione. Ciò che è comune a questo tipo di esperienze è il fatto che il flusso degli eventi non si arresta insieme alla presentazione della rottura, ma si evolve, in un modo o nell'altro: risulterebbe ancora più imbarazzante se, davanti a una piccola rottura come quella descritta, ci mettessimo a gridare in preda al panico o scappassimo via a gambe levate. Quello che si delinea è un panorama di possibilità, ognuna delle quali costituisce un percorso verso nuovi assetti dell'interazione: nel caso ci accorgessimo che il nostro interlocutore è timido, e per questo non ci ha stretto la mano come ci aspettavamo, potremo ricalibrare la nostra corporeità per metterlo a proprio agio, magari allontanandoci o facendo finta di non aver allungato la mano verso di lui; potremo scherzarci sopra, o fare finta di stare afferrando qualcosa dietro di lui. Quelle elencate sono solo alcune delle strategie di compensazione derivanti dalla necessità di controbilanciare un *breakdown* attraverso la stabilizzazione di un nuovo piano di significato: nella terminologia coniata da Varela, un altro *micromondo*.

In other words, in the breakdown before the next microworld shows up there is a myriad of possibilities available until, out of the constraints of the situation and the recurrence of history, a single one is selected. The fast dynamics is the neural correlate of the autonomous constitution of a cognitive agent. (Varela 1992b: 107)

I *micromondi* a cui Varela fa riferimento sono definibili come delle strutture dinamiche di significati che vengono enattivate in corrispondenza alla definizione della situazione interazionale. Il modo in cui è possibile riconoscere un micromondo è attraverso l'identificazione di ciò che, in una determinata circostanza, è pronto all'azione: una serie di *affordance*, quindi, che si propongono nella costituzione di un ambiente significativo sotto certi aspetti, rendendone virtuali degli altri, e definendo i confini delle possibilità interpretative e di azione al suo interno. Ciò che si configura è l'emersione di un pattern saliente di caratteristiche ambientali che si accordano alle possibilità di azione eseguibili in quella circostanza, in quel micromondo. In questo contesto, il *breakdown* si definisce come i) l'intervallo temporale, cognitivo e percettivo tra un micromondo e l'altro e, spesso, come ii) il motivo di passaggio da un micromondo all'altro, innescato dalla destabilizzazione del micromondo corrente.

L'analisi che Varela fa seguire alla identificazione di questa prospettiva dell'agente sul mondo e del repentino cambiamento che si realizza nel mondo dell'agente in riferimento alle sue possibilità di azione al suo interno ci consente di enucleare alcune delle caratteristiche da tenere in considerazione nell'analisi dei *breakdown*:

- i) Le rotture come quella descritta da Varela o quella che si delinea davanti a una stretta di mano mancata non sono eccezioni nella vita quotidiana. Sono, anzi, una regolarità: nel persistente e continuo flusso di senso che emerge dalla pragmatica dell'azione in un ambiente significativo, l'agente si trova costantemente a fare i conti con imprevisti, interruzioni, incertezze, vulnerabilità, iper e ipo-salienze che vanno riconfigurate e aspetti della materialità dell'ambiente da scoprire, che a volte sorprendono l'agente nel corso stesso del processo di manipolazione e gestione del significato. In sostanza, il processo di passaggio tra micromondi ed esperienza del *breakdown* è quello che costituisce il *sense-making* dell'organismo come attività ininterrotta di produzione di senso, «the constant bringing forth of signification [...] we may describe as a permanent lack in the living» (Varela 1991a: 8): la precarietà di un sistema che scopre, al contempo, sé stesso, le sue possibilità di azione e l'ambiente che lo ospita. Nel caso dell'agente umano, questi aspetti prendono anche la forma del pensiero cosciente, della capacità di ritornare, anche linguisticamente, sulle proprie azioni e sui propri pensieri, elaborando modalità di recupero della rottura che si disseminano su tanti piani pragmatici quante sono le capacità di *sense-making* che l'agente padroneggia. I *breakdown* sono quindi una costante del rapporto con il mondo, dal momento che puntellano il sentiero della cognizione incarnata modellandola attraverso i tentativi pragmatici dell'organizzazione locale del senso.

We always operate in some kind of immediacy of a given situation: our lived world is so ready-at-hand that we don't have any deliberateness about what is and how we inhabit it. [...] We have readiness-to-action which is proper to every specific lived situation. New modes of behaving and the transitions or punctuations between them correspond to mini (or macro) breakdowns we experience constantly. (Varela 1992b: 101)

- ii) La costituzione di un micromondo ha un carattere *ricorsivo*, o meglio si basa sull'esperienza che l'agente ha di situazioni simili in cui micromondi specifici sono stati enattivati in corrispondenza di *breakdown* interazionali. Nelle parole di Varela,

[...] being capable of appropriate action is, in some important sense, a way in which we embody a stream of recurrent microworld transitions. [...] In other words, microworlds/identities are historically constituted. But the pervasive mode of living consists of the *already* constituted microworlds which compose our identities» (Varela 1992b: 101).

In sostanza, la costituzione di un *micromondo*, ovvero una specifica configurazione di possibilità di interazioni che emergono all'interno di un ambiente, avviene in modo ricorsivo e sulla base della storia interazionale dell'agente, la cui attività consiste nel costruire e ricostruire tali micromondi sulla base dell'esperienza pregressa di situazioni simili. Ciò che accade è che ogni volta che un agente agisce all'interno del proprio ambiente e individua o percepisce un *breakdown*, enattiva una specifica configurazione interazionale per far fronte alla situazione. Questo implica anche che, secondo l'approccio enattivista, tali micromondi non sono entità statiche, ma configurazioni che si evolvono storicamente attraverso l'esperienza di transizioni ricorrenti. In questo senso, Varela sottolinea che la capacità di compiere azioni appropriate all'interno di una di queste configurazioni consiste in una manifestazione di come gli agenti incarnano un flusso continuo di passaggi consolidati e considerati efficaci tra micromondi. In altre parole, l'identità e il comportamento di un agente emergono dall'interazione storica con una serie di micromondi che si sono consolidati passando attraverso la prova dell'efficacia pragmatica. Ciò implica che la nostra esistenza quotidiana e le nostre capacità di interagire in modo efficace nell'ambiente non sono costituite da risposte a eventi singoli, ma piuttosto da un processo continuo di ricostituzione di micromondi basati sulle esperienze passate e insieme ad altri agenti immersi nel nostro stesso ambiente.

- iii) La storia dell'agente non è l'unico fattore che condiziona l'emersione di un micromondo. In questo processo, l'altro elemento fondamentale da considerare è l'immediatezza dell'interazione, di ciò che è presente al momento e che risulta rilevante per l'agente: un movimento, dunque, che non va solamente dalla mente al mondo, ma che si configura come il risultato di una congerie dinamica in cui vengono messe a disposizione possibilità reali e virtuali di azione e significazione. L'interazione stessa, quindi, influenza l'emersione del micromondo in quanto condizionamento materiale e pragmatico delle possibilità di azione.

At each such breakdown the manner in which the cognitive agent will next be constituted is neither externally decided nor simply planned ahead. It is a matter of *commonsensical emergence*, of an autonomous configuration of an appropriate stance. (Varela 1992b: 102)

Il movimento attraverso cui il micromondo si costituisce, dunque, si sviluppa a partire dalla configurazione del *breakdown* in interazione, dalla storia dell'agente all'interno di una situazione *embodied* simile fatta di circostanze possibili e impedimenti dell'azione, e dalla immediatezza dell'esperienza dell'agente, il quale si trova implicato nella contingenza di rispondere alle rotture e continuare a produrre significati.

I *breakdown*, in sostanza, si configurano come momenti in cui si verifica una transizione tra micromondi che implicano diverse possibilità e capacità di azione, diversi allestimenti di salienze percettive, diverse modalità di co-costruzione del senso. Come operazione preliminare, possiamo subito raffrontare la considerazione della rottura per come concettualizzata da Varela al ruolo che il momento di realizzazione dell'interrogatore umano ha nel test di Turing.

Nel classico gioco dell'imitazione *à la* Turing, il focus è sulla presenza o sull'assenza della rottura: il momento in cui avviene il riconoscimento del fallimento da parte della macchina di riprodurre la struttura dello scambio sancisce la vittoria dell'umano e la sconfitta della macchina. L'assenza di rotture, d'altra parte, segnerebbe un traguardo da parte della macchina, che si configurerebbe come agente cognitivo al pari dell'essere umano. In questo quadro, quindi, la rottura non è considerata come fattore insito alle logiche dell'interazione, ma come elemento che decreta la fine dei giochi.

La rottura dell'interazione è invece concettualizzata in maniera completamente diversa da Varela, in cui il gioco si sviluppa su un terreno sempre arrischiato, e include la rottura e il suo recupero come una sua parte costitutiva. In questa ottica, l'interazione è il terreno onnipresente del gioco della vita, che si deve adeguare a tutti gli imprevisti che si presentano, anche quando nemmeno lontanamente attesi o molto destabilizzanti. In questo senso, nel gioco dell'interazione è importante capire come la rottura avviene, se ci sono strategie e micromondi da enattivare che costituiscono scorciatoie di stabilizzazione del senso, le modalità in cui il *breakdown* si manifesta in modo da prevenire imprevisti e minimizzare lo sforzo interpretativo. I tentativi di recupero della rottura, nel gioco dell'imitazione, non sono opzionali, ma ciò che naturalmente segue dalla sua comparsa: si continua a giocare, quindi, cercando di recuperare fino al prossimo imprevisto.

Once a behavioral stance is selected or a microworld is brought forth, we can more clearly analyze its mode of operation and its optimal strategy. In fact, the key to autonomy is that a living system, out of its own resources, find its way into the next moment by acting appropriately. And it is the breakdowns, the hinges that articulate microworlds, that are the source of the autonomous and creative side of living cognition. Such commonsense, then, needs to be examined at a microscale: at the moments where it actualizes *during breakdowns*, the birthplace of the concrete. (Varela 1992b: 102)

Pertanto, per analizzare cosa avviene nell'interazione tra umani e robot nel suo complesso, ciò che si rende necessario è un'analisi di uno dei *breakdown* più significativi e complessi all'interno del campo degli studi sulla HRI, un fenomeno che è divenuto nel tempo una delle sfide più complesse e dibattute in questo ambito. Il fenomeno in questione è l'Uncanny Valley Effect (UVE), una manifestazione di disagio nell'essere umano di fronte al robot umanoide che si colloca al confine tra la percezione del familiare e dell'inquietante, e che continua a sfuggire a una comprensione stabile sia nelle scienze cognitive, sia negli studi dell'interazione tra umani e robot.



La prima formulazione dell'UVE risale agli anni '70, frutto dell'intuizione quasi profetica<sup>165</sup> del robotico giapponese Masahiro Mori, e il fenomeno è stato da allora oggetto di indagini sempre più articolate e controverse, sviluppate aderendo a diverse teorie sulla cognizione, sulla percezione e sulle possibili conseguenze che l'UVE nella HRI. Il suo carattere elusivo ha spinto gli studiosi a confrontarsi con interrogativi che mettono in relazione l'aspetto dell'artefatto robotico, le qualità che lo avvicinano all'umano e le la confidenza dimostrata da quest'ultimo di navigare interazioni con oggetti dalla particolare capacità sintetica, che ancora oggi, però, rimangono aperti: che cosa determina esattamente l'insorgere di questa sensazione di disagio nell'essere umano di fronte a robot che presentano caratteristiche *troppo* simili all'umano? Quali sono le conseguenze di questo disagio? E, soprattutto, come si può fornire una base sperimentale solida che permetta di spiegare ciò che accade a livello percettivo e comportamentale?

Nessuna delle molteplici definizioni proposte in letteratura è riuscita a ottenere un consenso unanime né sulla natura del fenomeno, né sulle sue cause profonde. L'UVE sembra, infatti, manifestarsi in una pluralità di forme, sfuggendo a una categorizzazione precisa e trovando antecedenti storici e culturali che ampliano la sua portata. In tale contesto, comprendere l'Uncanny Valley Effect diventa un nodo centrale per l'ottimizzazione delle interazioni tra esseri umani e robot, facendo di questo *breakdown* uno dei problemi più stimolanti e complessi di tutto il settore dell'HRI.

## 5.2 L'Uncanny Valley Effect

### 4.4.5 *L'effetto perturbante: identificazione di uno strano fenomeno e delle sue possibili cause*

L'*Uncanny Valley Effect* è un fenomeno concettualizzato per la prima volta dal robotico giapponese Masahiro Mori (Mori 1970). Inizialmente, la definizione è comparsa in un articolo pubblicato sulla rivista giapponese *Energy*, portando alla luce, senza evidenza scientifica a supporto ma sulla base dell'esperienza personale dello scienziato, un fenomeno che Mori riteneva tipico dell'interazione tra umani e robot: una sensazione negativa, inquietante e di disagio che mette l'osservatore umano in una situazione spiacevole nei confronti di alcuni robot. Secondo Mori, il fattore scatenante di questo fenomeno consiste nella somiglianza all'umano mostrata dal robot in questione: artefatti umanoidi come giocattoli, pupazzi o dispositivi tecnologici con fattezze umane stilizzate sembrano influire positivamente sulla qualità dell'interazione, incentivando un trattamento

---

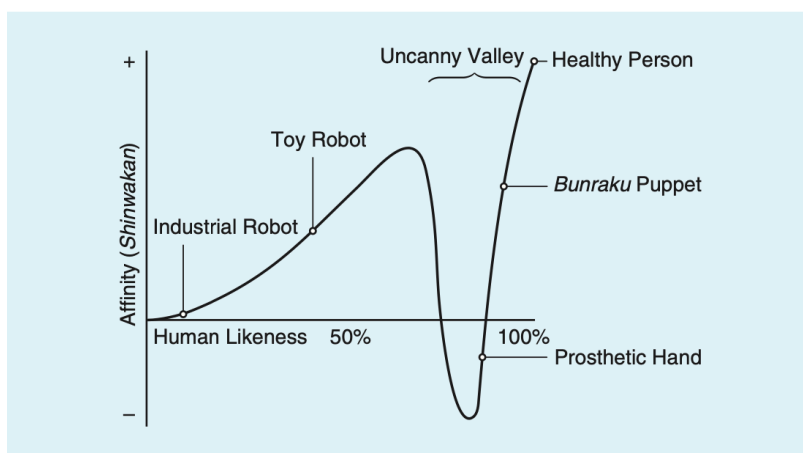
<sup>165</sup> Il carattere profetico è legato al fatto che, al momento della formulazione dell'ipotesi da parte di Mori, non fosse stato prodotto alcun esemplare di robot altamente antropomorfo. L'intuizione, quindi, si basa su un parallelo originale con altre forme di costruzione e invenzione realizzate dall'umano, come quelle delle figure tipiche del teatro giapponese e della scultura, oltre che sulle sporadiche realizzazioni di artefatti antropomorfici (come arti e parti del corpo) fino a quel momento effettivamente realizzate (cfr. Mori 1970).

da parte dell'umano che ricorda quello riservato a bambini o altri agenti sociali; improvvisamente, però, con artefatti che somigliano molto di più agli umani rispetto a quelli appena citati, come androidi, bambole di eccezionale fattura o altre riproduzioni fedeli della sembianza dell'essere umano, ciò che assale molti degli individui di fronte a loro è una sensazione ambigua, un misto di sorpresa, paura, disagio, disgusto e repulsione, che condiziona fortemente in negativo la possibilità di interazione.

Secondo Mori, questo fenomeno può essere descritto correlando in maniera non lineare la dimensione quantitativa del grado di somiglianza umana dell'artefatto robotico ( $x$ ) e la dimensione qualitativa della familiarità ( $y$ ) che caratterizza il modo in cui l'agente umano interagisce con esso. Come lo stesso Mori sottolinea,

An example of a function that does not increase continuously is climbing a mountain—the relation between the distance ( $x$ ) traveled by a hiker toward the summit and the hiker's altitude ( $y$ )—owing to the intervening hills and valleys. I have noticed that, in climbing toward the goal of making robots appear like a human, our affinity for them increases until we come to a valley, which I call the *uncanny valley*. (Mori et al. 2012: 98)

Pertanto, l'espressione *Uncanny Valley* (UV) si riferisce a una curva discendente in un grafico che rappresenta la relazione tra la qualità dell'interazione e il grado di somiglianza umana del robot. A partire da una somiglianza minima tra l'aspetto del robot e quello umano, in cui i robot presentano solo alcuni tratti stilizzati che ricordano l'umano in qualche modo, il grado di qualità dell'interazione aumenta, favorendo l'affidabilità e i meccanismi antropomorfici tipici della cognizione umana. Questo accade fino a quando la curva raggiunge un punto di catastrofe<sup>166</sup>, in cui, al crescere della somiglianza, la sensazione diventa inquietante e la qualità dell'interazione decade, sprofondando nella valle disegnata dal grafico (Figura 5.1).



<sup>166</sup> Un punto di catastrofe è una discontinuità puntuale e non lineare che si verifica in un sistema dinamico quando un cambiamento graduale nei parametri di controllo porta a una transizione improvvisa nello stato del sistema. In prossimità di questo punto critico, piccole variazioni di uno o più parametri possono causare un cambiamento sproporzionato nella risposta del sistema, determinando un passaggio repentino da uno stato di equilibrio a un altro (cfr. Thom 1975).

Fig. 5.1. «The graph depicts the uncanny valley, the proposed relation between the human likeness of an entity, and the perceiver's affinity for it. [Translators' note: *Bunraku* is a traditional Japanese form of musical puppet theater dating to the 17th century. The puppets range in size but are typically a meter in height, dressed in elaborate costumes, and controlled by three puppeteers obscured only by their black robes] » (Mori et al. 2012: 99)

È interessante sottolineare che Mori nota sin da subito che il fenomeno potrebbe essere condizionato dalla capacità del robot di muoversi autonomamente: l'effetto sarebbe meno potente se l'artefatto è statico, mentre i contesti dinamici «steepen the slopes of the uncanny valley» (Mori et al. 2012: 99) (Fig. 5.2).

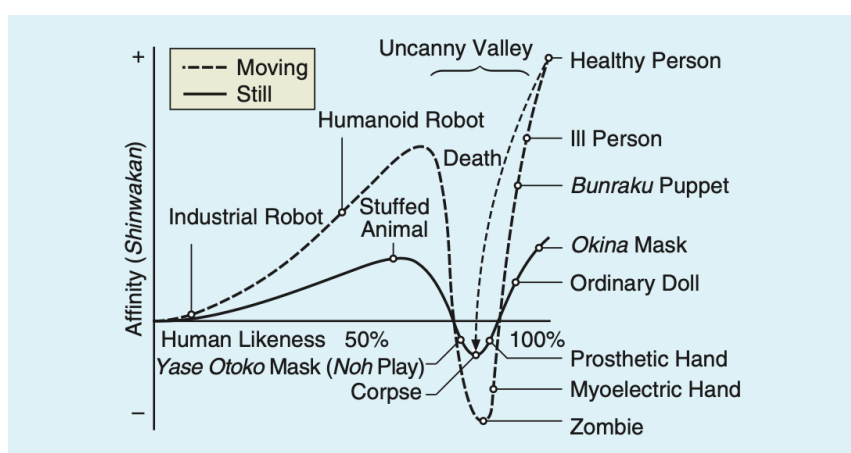


Fig. 5.2. «The presence of movement steepens the slopes of the uncanny valley. The arrow's path represents the sudden death of a healthy person. [Translators' note: *Noh* is a traditional Japanese form of musical theater dating to the 14th century in which actors commonly wear masks. The *yase otoko* mask bears the face of an emaciated man and represents a ghost from hell. The *okina mask* represents an old man]» (Mori et al. 2012: 99)

Il suggerimento che Mori inserisce come conclusione all'articolo del 1970 consiste nel tentativo di bypassare l'UV progettando robot che mostrino un grado moderato di somiglianza all'umano, consigliando agli ingegneri robotici di concentrarsi sul raggiungimento del primo picco – e non del secondo – nella curva rappresentativa della relazione tra *humanlikeness* e familiarità percepita.

Di per sé, l'articolo non ha sollevato molte attenzioni negli anni immediatamente successivi alla sua pubblicazione, tanto che rimase pressoché sconosciuto fino al 2005, quando fu tradotto in inglese da Karl MacDorman. A partire da quell'anno, l'attenzione verso il fenomeno dell'UV inizia a crescere: l'interesse verso il fenomeno si allarga al di fuori delle discipline immediatamente collegate alla realizzazione di robot umanoidi, arrivando a interessare coloro che volevano indagare le cause, le conseguenze e la struttura di fenomeni concernenti la dinamica interazionale con i robot. Soprattutto la psicologia, quindi, ha contribuito alla formalizzazione e alla diffusione del fenomeno

nei primi tempi dopo la traduzione dell'articolo (cfr. MacDorman & Ishiguro 2006; Moore 2012; Yin et al. 2021; Rosenthal-von der Pütten & Weiss 2015; Stein & Ohler 2017; Mara et al. 2022; Cheetham et al. 2015; Seyama & Nagayama 2007; MacDorman & Chattopadhyay 2016), sviluppando anche diverse riflessioni sulle possibili cause dell'esperienza inquietante e allacciando legami con fonti precedenti, come Freud (Tondur 2015; Benjamin & Heine 2023) e Jentsch (Stein & Ohler 2017; MacDorman & Entezari 2015) (cfr. Schwind 2015).

A nutrire ulteriormente l'interesse verso questo nuovo fenomeno ha avuto un ruolo anche una serie di modifiche e ulteriori sviluppi che lo stesso Mori ha apportato sia nella traduzione dell'articolo originale (Mori 1970), sia in articoli successivi (Mori et al. 2012). In primo luogo, Mori chiarisce la terminologia utilizzata nel 1970, soprattutto in riferimento alla dimensione relativa alla qualità dell'interazione. In particolare, Mori sostituisce il termine *familiarity* con *affinity*, dal momento che il primo avrebbe potuto suggerire significati legati a livelli di novità o confidenza, discostandosi dal termine giapponese originale *shinwa-kan*. Tuttavia, anche questa sostituzione non ha eliminato del tutto i fraintendimenti e il termine continua a generare ambiguità sia nella ricerca filosofica, sia negli studi sperimentali, tanto che studiosi provenienti da diversi settori disciplinari stanno cercando di individuare termini più adeguati a rendere il concetto originario giapponese, che sembra non avere una corrispondenza precisa in altre lingue (Pollick 2009; Bartneck et al. 2009; Misselhorn 2009).

In secondo luogo, alle possibili spiegazioni del fenomeno che aveva proposto nell'articolo originale, il robotico giapponese aggiunge una serie di plausibili cause legate all'evoluzione dell'essere umano e al suo legame sociale con i conspecifici. Da una parte, Mori spiega il disagio suscitato nell'UVE collegando la sensazione inquietante al fatto che questi robot, ricordando corpi privi di vita, possano sembrare cadaveri agli occhi degli esseri umani (Mori et al. 2012: 100). D'altra parte, Mori suggerisce che l'inquietudine potrebbe essere data dal fatto che gli umani, interagendo con i robot, possano in questo modo scoprire il "trucco" dietro la loro *apparente* somiglianza all'umano e sentirsi quindi ingannati dal fatto che non riescano a riprodurre fedelmente anche il suo comportamento: in sostanza, il senso di disagio che accompagna l'aumento della somiglianza all'umano potrebbe derivare dalla percezione di questo aspetto ingannevole dei robot. In conclusione, Mori ipotizza anche che alcune entità potrebbero riuscire ad oltrepassare la *valley*, facendo l'esempio di alcune statue buddhiste che appaiono avere un portamento elegante e una calma serafica, ovvero rappresentazioni idealizzate dell'essere umano perfetto: qualcosa che risulta "più umano dell'umano" (Zlotowski et al. 2013).

Pur non avendo fornito linee guida precise per testare o investigare l'ipotesi dell'UV – o forse proprio a causa di questa mancanza – il fenomeno divenne rapidamente un tema centrale negli studi sull'interazione umano-robot, attirando oltre agli interessi degli psicologi anche quelli dei designer

(Tinwell 2014; Brenton et al. 2007; Schwind et al. 2018; Bakan 2021), dei filosofi (Angelucci et al. 2016; Gahrn-Andersen 2022; Misselhorn 2009; Johnston et al. 2020; Fuchs 2021, 2024) e degli scienziati cognitivi (Wang et al. 2015; Wang & Rochat 2017; MacDorman & Ishiguro 2006; Cheetham (Ed) 2018; MacDorman 2005; Ho & MacDorman 2017)<sup>167</sup>.

Possiamo tentare di raccogliere le indagini sull'UVE provenienti da vari campi disciplinare negli ultimi venti anni guardando a come diversi studi hanno tentato di rispondere a tre delle domande centrali riguardo al fenomeno:

- *L'UVE esiste? E se sì, è un fenomeno universale nell'esperienza umana di fronte al robot?*

Dal momento che Mori non ha corredato le sue intuizioni con alcuna evidenza scientifica del fenomeno dell'UV, una delle principali indagini che si è delineata subito dopo la traduzione dell'articolo consiste nella sua validazione sperimentale, con lo scopo di sancire la sua effettiva esistenza all'interno della sfera dell'esperienza e della psicologia umana. In generale, si può riscontrare un riconoscimento diffuso del fenomeno in termini scientifici fin dagli inizi, a rafforzare l'ipotesi che l'effetto sia generato dalla relazione tra familiarità dell'interazione per l'essere umano e somiglianza all'umano del robot, come predetto da Mori (MacDorman & Ishiguro 2006; Seyama & Nagayama 2007). In seguito, l'ipotesi è stata provata in termini sperimentali per esseri umani adulti (Burleigh et al. 2013; Ho et al. 2008; MacDorman et al. 2009; Mitchell et al. 2011; Seyama & Nagayama, 2007; Tinwell et al. 2011, 2015; Wang & Rochat 2017), per bambini<sup>168</sup> (Tinwell & Sloan 2014; Brink et al. 2019; Strait et al. 2019; Chien et al. 2024), per neonati (Lewkowicz & Ghazanfar 2012; Matsuda et al. 2012) e per primati non umani (Steckenfinger & Ghazanfar 2009; Woods & Hare 2019). Tuttavia, alcuni studi testimoniano della possibilità che, a volte, l'effetto non venga registrato (Pollick 2010; Sansoni et al. 2015; Stix 2008; Cheetham et al. 2014; cfr. Brenton et al. 2005;), e ciò viene motivato con una serie di ragioni spesso relative all'apparato sperimentale (Kätsyri et al. 2015) o alle inconsistenze teoriche di fondo (Wang et al. 2015).

Uno dei problemi legati alla individuazione e definizione del fenomeno è riscontrabile nella scelta originaria di Mori di chiamare l'UV con il termine giapponese 'Bukimi' (不気味), in seguito tradotto con l'inglese 'uncanny'. Il termine sembra non avere una perfetta traduzione in altre lingue e, a detta di uno degli studiosi che più si è occupato del tema, la traduzione più corretta del termine sarebbe «valley of eeriness» (MacDorman citato in Hsu 2012), ovvero valle dell'inquietudine, e non la sua

---

<sup>167</sup> Per una *review* degli approcci coinvolti nello studio del fenomeno dell'UVE, si veda: Zhang et al. (2020); Wang et al. (2015); Vaitonyté et al. (2023).

<sup>168</sup> È interessante notare che alcuni studi riscontrano un'assenza di UVE nei bambini diagnosticati con la sindrome dello spettro autistico (Feng et al 2018).

versione più diffusa *uncanny valley*<sup>169</sup>. Tuttavia, questa seconda versione si è affermata nel tempo anche perché apriva il campo a una serie di studi già consolidati in psicologia e psicanalisi sul tema del *perturbante*. Il legame istituito con questa tradizione ha contribuito in modo sostanziale a modellare il significato del concetto, trasferendo le connotazioni sviluppate per il perturbante al fenomeno di disagio di fronte al robot. Il legame istituito, ovviamente, affonda le sue radici nella formulazione psicanalitica del perturbante: già Freud (1919) aveva sottoposto il fenomeno ad analisi in uno dei suoi saggi più celebri, in cui parlava di una sensazione di paura e inquietudine associata alla novità e alla non familiarità del rapporto con oggetti o porzioni del mondo esterno. Il nodo cruciale, poi, è stato ricoperto dal riferimento di Freud all'opera *The Sandman* di Hoffman (1817), in cui sono numerosi i richiami a una statua dalle fattezze estremamente umane. Il primo, tuttavia, a parlare di perturbante in relazione a un artefatto umanoide è Ernst Jentsch, psichiatra tedesco che, facendo riferimento alla stessa opera letteraria, fornisce la prima spiegazione del fenomeno collegata all'incertezza decisionale: parlando dell'opera di Hoffman, infatti, Jentsch afferma che la sensazione di disagio è data dalla difficoltà di decidere se uno dei personaggi principali del racconto, ovvero l'automa Olimpia, potesse essere realmente una persona (cfr. Kang 2011). Senza dubbio, quindi, il panorama concettuale che definisce il fenomeno dell'UV è ad oggi il frutto di numerose ibridazioni, sia transculturali, sia storiche, sia inerenti al dominio psicologico e psicanalitico, in cui il legame con gli studi classici sul tema del perturbante ha ricoperto un ruolo fondamentale: come sottolineato da Wang e colleghi (2015: 401), «because 'uncanny' is also close in meaning, although not equivalent, to the German term 'Unheimliche', it is therefore also linked to its aforementioned interpretations in the German literature».

In ogni caso, anche se sono riscontrabili evidenze in cui l'UVE non viene registrato o si fatica a uniformare l'oggetto di indagine sotto un'unica denominazione comune a più sguardi, il fenomeno sembra essere effettivamente rilevante nella maggior parte dei casi in cui un essere umano si trova di fronte a un robot altamente *human-like* (Koopman 2019; Wang et al. 2015; Wang & Rochat 2017; Mathur & Reichling 2016; Ishiguro 2006, 2007). La domanda, quindi, da essere circa l'esistenza del fenomeno, si sposta sulle modalità attraverso cui l'UVE è stato formalizzato e sulle dimensioni che si credano possano descriverla.

#### - *Il grafico disegnato da Mori descrive L'UVE?*

Un cospicuo numero di studi riporta l'evidenza di un fenomeno identificabile con l'UVE nell'esperienza di fronte al robot, ma non nei termini descritti da Mori. Alcuni studi, infatti,

---

<sup>169</sup> Questa traduzione appare per la prima volta in Reichard (1978) e da quel momento è stata considerata la traduzione ufficiale del termine giapponese.

testimoniano della difficoltà nel trovare una correlazione tra le dimensioni ipotizzate da Mori alla base dell'effetto (Brenton et al. 2005; Hanson et al. 2005; MacDorman 2006; Pollick 2009; Poliakoff, Beach et al. 2013). Ad esempio, Pollick (2009) mette in discussione proprio i termini con cui l'UVE è descritta, avanzando anche dubbi su quella che fino ad allora era stata la modalità sperimentale di investigazione del fenomeno, ovvero quella della valutazione di *morphed image*<sup>170</sup>. Nel suo articolo, lo studioso cerca di fornire una definizione formale e operativa del fenomeno, ponendo l'accento sulla necessità che le ricerche future non si facciano condizionare dalle prime formulazioni del fenomeno<sup>171</sup>. Dello stesso avviso sono Brenton e colleghi (2005), che fanno notare la mancanza di una base sperimentale stabile e ipotizzano che il fenomeno potrebbe essere soggetto a molte influenze di fattori esterni, supposizione a seguito verificata per l'influenza dell'età (Tu et al. 2020), la cultura di appartenenza (Draude 2011; Ho et al. 2017) e altri fattori relativi all'esperienza individuale (MacDorman & Entezari 2015; Destephe et al. 2015).

Il dibattito sulla affidabilità delle dimensioni scelte da Mori si è sviluppato negli anni anche attraverso la proposta di diverse forme e figure che possono risultare dal grafico che mette in relazione affinità o familiarità e somiglianza all'umano. Ad esempio, nello studio di Bartneck e colleghi (2007) sulla *likeability* suscitata da diverse entità quali umani, robot umanoidi e androidi, i ricercatori propongono di rinominare il fenomeno passando da *uncanny valley* a *uncanny cliff*: un precipizio, dunque, in cui il secondo picco di familiarità dell'interazione – quello che nel grafico di Mori identificherebbe il momento massimo di risalita dalla valle – è più basso del primo. In sostanza, la maggiore familiarità nell'interazione si risconterebbe, secondo gli autori, con artefatti umanoidi che dimostrano una media somiglianza all'umano, come giocattoli e robot dai tratti fortemente stilizzati, e non nel secondo ramo ascendente della curva. Diversamente, Tinwell et al. (2011) propongono di sostituire l'immagine della valle con quella di un "muro", riscontrando l'impossibilità nello stato dell'arte sugli esperimenti per l'UVE di valicare il baratro dell'inquietudine. Un'ulteriore variazione del grafico che descrive il fenomeno è fornita da Kim e colleghi (2022), i quali hanno progettato un esperimento in cui sono state considerate risposte emotive sia positive sia negative a immagini a figura intera di robot. Hanno scoperto non una, ma due *uncanny valley*: la prima corrisponde a quella

---

<sup>170</sup> La tecnica delle *morphed image* è senza dubbio quella più adottata ancora oggi per testare l'UVE, soprattutto per validare teorie riguardo alla percezione categoriale (cfr. §5.2.1.2; 5.2.2). Questa tecnica sperimentale consiste nella creazione di una transizione graduale tra due immagini, spesso utilizzando software di elaborazione grafica per fondere caratteristiche visive di due entità distinte. Questo processo genera una serie di immagini intermedie che variano lungo un continuum, ad esempio tra due diverse espressioni facciali o tra un volto umano e un volto robotico. Lo scopo è studiare come i partecipanti percepiscano il cambiamento progressivo e identificare i punti di confine in cui avviene una transizione percettiva significativa (cfr. Sutherland et al. 2017). Nel caso dell'utilizzo per testare l'UVE, l'interesse si concentra nel passaggio da "non umano" a "umano", o viceversa (Cheetham et al. 2011; Mathur & Reichling 2016; MacDorman et al. 2009; Burleigh et al. 2013; Seyama & Nagayama 2007; Kätsyri et al. 2015).

<sup>171</sup> «[...] there is a dearth of empirical evidence on the topic and certainly no study that outlines essential properties that can be manipulated to navigate into and out of the uncanny valley». (Pollick 2009)

teorizzata da Mori, per robot umanoidi altamente realistici; la seconda emerge invece per robot con tratti moderatamente umanoidi. I ricercatori interpretano la creazione di questo bivio nell'esperienza davanti alle entità somiglianti all'umano associando rispettivamente le due valli a una risposta emotiva negativa (*bukimi*) e a una positiva (*shinwakan*) (cfr. anche Ho & MacDorman 2010).

Un'altra serie di esperimenti ha aderito alla teoria di Mori riguardo alla valle del perturbante, ma non alla sua ipotesi secondo cui la dinamicità e i movimenti degli artefatti ne amplificherebbero l'effetto (Piwek et al. 2014; Thompson et al. 2011), o hanno messo in dubbio la sua generalizzazione, sottolineando come tale fenomeno sembri verificarsi solo in circostanze specifiche (Kätsyri et al. 2015; Pollick 2009). Altri hanno messo in discussione il fatto che l'UVE si caratterizzi esclusivamente come esperienza negativa di fronte al robot, come già suggerito da Kim et al. (2022) e anticipato dalle ricerche di Hanson e colleghi (2005) sul valore estetico dell'UVE (cfr. § 5.2.1.1). L'evidenza più forte in questo senso proviene dallo studio di Cheetham e colleghi (2014), i quali hanno condotto tre esperimenti utilizzando le *morphed images* di volti in un continuum di stimoli da completamente artificiale a completamente umano. I test prevedevano che il soggetto decidesse se l'immagine-stimolo che gli veniva presentata fosse meglio descrivibile come artificiale o umana e hanno rilevato che la difficoltà nella discriminazione è minore per volti ambigui e avatar non ambigui, mentre aumenta per i volti umani non ambigui. Di conseguenza, hanno evidenziato che «ratings of familiarity show that faces associated with greatest category ambiguity do not show an uncanny-like effect» (Cheetham et al. 2014: 13). Al contrario, sostengono che, poiché «negatively valenced ratings increased across the tested stimulus conditions with increasing distance from the human end of the continua» (ibid.), e la difficoltà nella discriminazione percettiva risulta meglio associata a sentimenti positivi, l'*uncanny valley* dovrebbe essere sostituita con la metafora della *happy valley*.

A mettere ancora in dubbio la visione di Mori c'è una ulteriore serie di studi che indaga l'appropriatezza dei termini scelti per descrivere il fenomeno, soprattutto in relazione alla traduzione in inglese dal giapponese dei termini impiegati. L'attenzione è stata principalmente diretta alla dimensione qualitativa del grafico espressa dall'asse y, in giapponese denominata *shinwa-kan*, in inglese tradotta con il termine "familiarity"<sup>172</sup> (Bartneck et al. 2007, 2009; Ho & MacDorman 2010). Uno dei primi studi ad avviare l'indagine sulla appropriatezza della traduzione è fornito da Bartneck e colleghi (2007), che mettono in luce delle possibilità di perdita del significato del termine originale nella traduzione. Gli autori sottolineano che il termine *shinwa-kan* non è di uso comune in giapponese, che non ci sono attestazioni di traduzioni dirette all'inglese e che la scelta di tradurre il termine con

---

<sup>172</sup> Come riportato da Kätsyri et al. (2015: 3): «The negative term *bukimi* translates quite unequivocally into eeriness (Ho & MacDorman 2010), although other similar terms such as creepiness and strangeness have also been used (cf. Ho et al. 2008). In contrast, the positive term *shinwakan* is an unconventional Japanese word, which does not have a direct equivalent in English».



l'inglese "familiarity" comporta non pochi problemi dal punto di vista della sua operativizzazione, dal momento che dipende fortemente dalla sua concettualizzazione in relazione alla cultura di appartenenza del soggetto e della sua esperienza, oltre che dalle possibilità di precedenti interazioni con i robot (Bartneck et al. 2007: 368). In un paper successivo, Bartneck e colleghi (2009) vanno ancora più a fondo alla questione, considerando i distinti nuclei semantici convocati dalla parola *shinwa-kan*:

The Daijirin Dictionary (second edition) defines *shinwa* as "mutually be friendly" or "having similar mind". *Kan* is being translated as "the sense of". Given the different structure of Japanese and English, not perfect translation is possible, but "familiarity" appears to be the less suitable translation compared to "affinity" and in particular to "likeability". After an extensive discussion with native English and Japanese speakers we therefore decided to translate "*shinwa-kan*" as "likeability". (Bartneck et al. 2009: 270)

Nelle conclusioni, gli autori ammettono che la questione si mostra tutt'altro che risolta, aprendo alla possibilità di conservare il termine nella sua versione originale giapponese come in alcuni casi di riconosciuta difficoltà di traduzione<sup>173</sup>, che tuttavia potrebbe portare ad una ancora maggiore difficoltà di resa operativa del termine nei contesti sperimentali. Come sottolineato in precedenza, la più recente traduzione dell'articolo in cui l'UVE è stata concettualizzata (Mori et al. 2012) riporta la traduzione del termine con 'affinity' che, se considerato nella sua negazione, è reputato essere l'alternativa migliore per descrivere la sensazione negativa innescata nella visione di un robot altamente antropomorfo.

In aggiunta a questi studi sulla traduzione dei termini, altri hanno invece cercato di trovare un modo migliore per descrivere il fenomeno in oggetto. Ad esempio, Misselhorn (2009) basa la sua indagine filosofica dell'UV sul fatto che la repulsione suscitata dall'effetto possa essere ricondotta a una mancanza di empatia, causata dalla percezione di un agente apparentemente vivo e dotato di anima, che tuttavia si rivela essere «cold and dead» (2009: 357). D'altra parte, Ho et al. (2010) propongono di integrare il concetto di 'warmth', considerandolo «virtually synonymous» (2010: 1508) di 'likeability', per misurare meglio gli effetti interpersonali e sociali dell'UVE. Nelle conclusioni, poi, gli autori dello studio suggeriscono che il fenomeno possa essere molto più sfaccettato di quanto inizialmente ipotizzato, invitando a considerare una multidimensionalità delle emozioni e delle esperienze che comprende paura, disgusto, nervosismo, avversione e sorpresa. Redstone (2013), d'altra parte, si concentra solamente sulla dimensione ansiogena del fenomeno, interpretando l'inquietudine provocata dall'UVE come una forma di angoscia, una paura

---

<sup>173</sup> Il termine portato ad esempio dagli autori è quello del *kansei engineering*, traducibile con 'ingegneria emotiva', che tuttavia non rende nel migliore dei modi il significato originale: tale branca dell'ingegneria, infatti, si riferisce a una metodologia di sviluppo del prodotto industriale che tiene in considerazione le emozioni, le sensazioni e le impressioni del cliente finale per la strutturazione dei parametri di design (cfr. Ishihara et al. 2008).

caratterizzata da incertezza o ambiguità riguardo a ciò che potrebbe influire sull'identità o sull'esistenza personale.

A fronte di questa serie di riflessioni che mettono in dubbio la descrizione del fenomeno, l'indagine sul disagio davanti al robot umanoide, nonostante la possibile arbitrarietà dei termini utilizzati per inquadrarla, sembra avere radici antiche e aprirsi anche a entità diverse dalle più recenti realizzazioni robotiche: il sentimento inquietante di essere di fronte ad artefatti dalle sembianze umane, infatti, è attestato ben prima dell'avvento dei robot umanoidi, come espresso nelle testimonianze di Jentsch (1906) e Freud (1919) agli inizi del XX secolo. Approfondendo ulteriormente, è possibile rintracciare descrizioni di fenomeni simili nelle statue – come il Golem (Sini 2019; Kang 2011) –, nelle bambole e nei clown (Lay 2015; Seyama e Nagayama 2017), ma anche nell'esperienza di fronte a pazienti con disturbi psicopatologici o psichiatrici (Fuchs 2019). L'UVE, in sostanza, sembra essere un fenomeno ampio, che ammette numerose sfaccettature e non limitato ai soli robot umanoidi o personaggi virtuali. Questa osservazione ci invita a sviluppare un approccio al fenomeno che non solo tenga conto della particolarità del robot in interazione, ma che istituisca paralleli con altri studi che si caratterizzano per la loro indagine a fenomeni legati alla sfera emotiva dei soggetti coinvolti.

- *Quali sono le cause dell'UVE?*

Sicuramente, questa è la domanda che ha raccolto più attenzioni da parte dei ricercatori sull'UVE, specialmente in ambito psicologico. In generale, come riportato nella totalità delle *review* sistematiche sul tema (MacDorman et al. 2009; Wang et al. 2015; Misselhorn 2009; Zhang et al. 2020; Kätsyri et al. 2015; Vaitonyté et al. 2023), è possibile raccogliere i tentativi di spiegazione del fenomeno in due gruppi: il primo riunisce le spiegazioni che fanno riferimento a processi cosiddetti *low-level*, ovvero relativi a capacità e meccanismi percettivi e alla sfera evolutiva dell'essere umano; l'altro gruppo, invece, raccoglie quelle spiegazioni che collegano il fenomeno a forme cognitive *high-level*, come la categorizzazione e la comprensione razionale dell'esperienza in questione. Affronteremo in ordine ognuna delle spiegazioni annoverate rispettivamente al livello percettivo e a livello cognitivo, allo scopo di fornire una panoramica adeguata delle cause considerate essere all'origine del fenomeno dell'UV. Possiamo, tuttavia, fornire un'osservazione preliminare, a partire proprio dal modo in cui le diverse teorie riguardo le cause e la struttura del fenomeno vengono suddivise: già dalla divisione tra spiegazioni percettive e cognitive, infatti, accolta dalla maggior parte degli studi, si riesce a comprendere quanto ancora lo sguardo alla mente e alla cognizione si componga a partire da una teoria classica delle scienze cognitive, vicina alle teorie funzionaliste e computazionaliste della mente e decisamente agli antipodi rispetto alla visione incarnata ed enattiva che stiamo cercando di proporre in questa tesi. Secondo tutti gli approcci che analizzeremo a breve,

la percezione e la cognizione sono due meccanismi da analizzare e considerare separatamente, la prima considerata il processo di raccolta e selezione degli stimoli esterni che la seconda deve analizzare e interpretare: anticipiamo che, ancora una volta, dunque, saremo costretti a notare che il corpo, l'esperienza incarnata e l'interazione viva sono fattori lasciati fuori dall'indagine e che, per queste teorie, tutto si svolge all'interno della mente dell'individuo osservatore, confinata nel suo cervello e praticamente insensibile ai cambiamenti che avvengono nell'ambiente in cui l'agente stesso è inserito. Abbiamo già visto come questo tipo di impostazione comporti delle notevoli conseguenze nell'analisi dell'antropomorfismo (§3.4.1) e dell'attribuzione di intenzionalità ai robot sociali (§4.3); ora ci concentreremo sulle conseguenze che questa comporta nella considerazione del fenomeno dell'UVE e dell'analisi dell'interazione tra umani e robot nel suo complesso.

A partire dalla divisione tra meccanismi percettivi e cognitivi, quindi, vediamo come le varie teorie ad oggi più diffuse nello stato dell'arte delle indagini sull'UVE spieghino il fenomeno dell'inquietudine di fronte ad artefatti altamente antropomorfi.

#### *4.4.5.1 Spiegazioni percettive*

Gli studi raccolti sotto la categoria di spiegazioni percettive descrivono l'UV come un «automatic, stimulus-driven, specialized processing that occurs early in perception» (MacDorman et al. 2009: 696; Wang et al. 2015: 394). In altre parole, le spiegazioni percettive cercano di spiegare l'UVE concentrandosi sulla risposta sensibile alla identificazione dello stimolo percettivo: il problema, secondo queste teorie, risiederebbe nell'individuazione di qualcosa che non va nella configurazione estetica dell'oggetto osservato, promotore dello stimolo.

Il fatto che l'UVE potesse avere origine a questo livello era già stato in qualche modo suggerito da Mori (1970), che aveva collegato l'emersione dell'effetto alla percezione sinestesica di una mano protesica: «One might say that the prosthetic hand has achieved a degree of resemblance to the human form. [...] However, once we realize that the hand that looked real at first sight is actually artificial, we experience an eerie sensation (Mori et al. 2012: 99), suggerendo inoltre che l'inquietudine sia il risultato di uno scollamento percettivo tra il realismo della mano sintetica e la strana sensazione nel toccarla o nell'immaginare di toccarla. In particolare, sotto questa categoria esplicativa ricadono tre teorie maggiori: la *pathogen avoidance hypothesis* (MacDorman et al. 2009; MacDorman & Ishiguro 2006), la *mortality salience hypothesis* (Ho et al. 2008; MacDorman & Ishiguro 2006) e la *evolutionary aesthetic hypothesis* (Hanson et al. 2005; MacDorman & Ishiguro 2006).

- *Pathogen Avoidance Hypothesis* (PAH). Questa teoria propone che volti di robot e altri artefatti dalle caratteristiche simili all'umano che evocano effetti di inquietudine siano per l'essere umano

indicativi di un alto rischio di malattie trasmissibili. In altre parole, l'incontro con creature che ricordano i nostri consimili ma che presentano caratteristiche che li fanno risultare strani innescerebbe una risposta evolutiva che implica il disgusto e la conseguente avvertita necessità di allontanarsi dall'entità in questione. Basata su alcune ricerche nel campo della psicologia evolutiva (Rozin & Fallon 1987), questa ipotesi riprende l'idea di Mori che l'UVE possa giocare un ruolo fondamentale nel processo di auto-preservazione della specie umana (Mori 1970) e viene proposta come una delle prime cause del fenomeno. Come riportato da Wang et al. (2015: 395), «because more humanlike entities are perceived to be more closely related to humans genetically, their imperfections elicit stronger aversion than do mechanical-looking entities». Reazioni fisiche come il disgusto (MacDorman & Entezari 2015; MacDorman & Ishiguro 2006; Wicker et al. 2003) e la paura (MacDorman et al. 2009) di fronte a un robot altamente umanoide, allora, sono state ricondotte a questo meccanismo di auto-preservazione che si innescerebbe qualora l'essere umano riconosca l'entità artificiale come potenzialmente pericolosa dal punto di vista della sua salute. In questo senso, alcuni studi hanno sottolineato che questa sensazione possa essere rilevata anche di fronte a soggetti con tratti psicopatici di personalità, che mostrerebbero un corrispettivo estetico al disturbo psichico nella conformazione fisica del soggetto considerato *uncanny* (Tinwell et al. 2013). Anche mettendo da parte la suggestione lombrosiana, c'è da sottolineare che l'ipotesi non risulta essere stata mai testata sperimentalmente (Wang et al. 2015; Wiese & Wiese 2020; Zhang et al. 2020), fattore che la fa risultare più una speculazione che una ipotesi operativa vera e propria.

- *Mortality Salience Hypothesis* (MSH). Legata alla PAH, la MSH sostiene che il disagio di fronte al robot è provocato non tanto dalla paura di essere contagiati da qualche malattia, ma dal fatto che l'entità artificiali ricordi all'umano la sua natura mortale. L'inquietudine, in questo caso, è sollecitata dal fatto che l'aspetto del robot o dell'androide potrebbe ricordare quello di un cadavere, un corpo senza vita con cui l'umano dovrebbe avere a che fare. Torna così ad essere rivalutata l'antica paura del *Doppelgänger* (cfr. §1.1-1.2), un'entità priva di anima che tuttavia ha la capacità di muoversi e forse non controllare la propria forza (Ho et al. 2008). I sostenitori di questa ipotesi (MacDorman 2005; Mac Dorman & Ishiguro 2006), appoggiandosi ad alcune teorie sociologiche ed antropologiche come la *Terror Management Theory* (TMT, Pyszczynski et al. 1999), leggono il fenomeno dell'UVE come una reazione spontanea alle repliche umanoidi, le quali sembrerebbero inquietanti poiché innescerebbero nell'umano un sistema di difesa che gestisce l'ansia per la morte, considerata profondamente radicata nella struttura dell'essere umano. In particolare, la TMT ipotizza che a pensieri riguardanti la morte corrispondano due reazioni principali: una strategia indiretta o a lungo termine di gestione della consapevolezza della morte, detta distale, che implica l'attaccamento a valori culturali, credenze religiose, ideali di successo o

identità collettive, ovvero una strategia che non affronta direttamente la paura della morte ma offre un senso di appartenenza e immortalità simbolica (ad esempio, l'idea di lasciare un'eredità o di essere ricordati); una strategia diretta, detta prossimale, che contempla risposte più immediate e consapevoli della minaccia della morte soprattutto nei momenti in cui si avverte una possibilità concreta di pericolo che innesci il tentativo di evitare situazioni pericolose, l'attenzione alla sicurezza personale, o la salvaguardia della propria salute. Alcuni ricercatori ipotizzano che l'UVE sia legata alle strategie distali, dal momento che l'esperienza di fronte al robot umanoide richiamerebbe valori simbolici come la vita, la sopravvivenza e un sentimento di pericolo non immediato<sup>174</sup> (MacDorman 2005). MacDorman e Ishiguro (2006) hanno provato a conferire validità scientifica all'ipotesi, allestendo un esperimento in cui mostrano ad alcuni soggetti l'immagine di un androide e quella di un essere umano, i quali in seguito hanno dovuto rispondere a una serie di domande poste in un questionario in cui veniva valutata la loro visione del mondo e il loro grado di protezione della propria salute. I risultati, anche se sono interpretati in favore dell'ipotesi della MSH, hanno tuttavia mostrato alcune incongruenze nelle modalità di misurazione dell'esperienza dei soggetti: alcuni hanno criticato l'utilizzo di misure esplicite, quando per misurare uno stato di paura o repulsione sembrerebbe più appropriato l'impiego di misure implicite dell'emotività e dello stato di agitazione dei soggetti (Wang et al. 2015: 395). Sebbene in misura minore rispetto alla PAH, anche in questo caso, dunque, vista oltretutto la vicinanza delle due ipotesi, le evidenze sperimentali non sembrano corroborare del tutto la validità della MSH, che rimane comunque un interessante sguardo sull'UVE dal punto di vista della filosofia dell'essere umano ma risulta di difficile integrazione nel panorama sperimentale correntemente adottato nella HRI.

- *Evolutionary Aesthetic Hypothesis* (EAH). Questa ipotesi è stata formulata per la prima volta da Hanson (2005), il quale ha proposto un legame tra l'inclinazione dell'essere umano a preferire alcune caratteristiche fisiche che segnalano una *fitness* evolutiva, come la fertilità o la salute, determinata da una pressione naturale, e il sentimento di disagio di fronte a una replica dell'essere umano. L'*uncanniness* sarebbe causata dalla bassissima attrattività dal punto di vista fisico nei confronti della replica umana, la quale non presenterebbe livelli accettabili di simmetria bilaterale, di qualità della pelle o tratti facciali e corporei nella media. Come riassunto da Geue (2021: 9),

---

<sup>174</sup> A ben vedere, tra i due meccanismi quello più pertinente nel caso dell'UV sembrerebbe quello prossimale, ma questa opzione non è presa in considerazione dagli studi che aderiscono alla teoria, dal momento che il soggetto sembrerebbe non mostrare, nella maggior parte dei casi, delle reazioni fisiche evidenti (cfr. MacDorman 2005; MacDorman & Ishiguro 2006).

The evolved cognitive and perceptual mechanisms seem to let an individual lacking reproductive fitness be perceived as unattractive (MacDorman et al. 2009). This in turn may produce a negative emotional response in order to avoid mating with this individual.

Hanson (et al. 2005) ha testato questa ipotesi somministrando ai soggetti dell'esperimento un questionario da compilare con le loro impressioni a seguito dell'osservazione di *morphed images* in cui all'estrema sinistra è rappresentato un robot e all'estrema destra un essere umano, cosicché la figura rappresentata si trasformasse da artificiale e quasi-naturale. I risultati dello studio mostrano che, se il soggetto valuta come inquietante le rappresentazioni che mostrano repliche umanoidi, la modificazione dell'immagine verso forme di apparenza più 'naturali' sembrano far sparire l'effetto *uncanny*. Come proposto dall'ipotesi, dunque, l'effetto di straniamento e inquietudine sembrerebbe causato da una repulsione nei confronti di tratti estetici non conformi a quelli umani o, in generale, che risultano disarmonici.

In generale, numerose sono state le critiche a questa impostazione sperimentale (Wang et al. 2015; Zhang et al. 2020), atte soprattutto a evidenziare l'inadeguatezza delle modalità di misurazione del test. Come reso chiaro da Wang et al. (2015: 395), «by simultaneously manipulating realism and attractiveness, Hanson introduced a potential confound and was unable to dissociate the effects of the two variables in determining human replicas' uncanniness».

Riassumendo, queste tre ipotesi legate alle spiegazioni percettive si basano principalmente su una capacità evolutiva dell'essere umano di individuare compagni tra i consimili e di evitare incontri che potrebbero risultare pericolosi o svantaggiosi per sé stessi. Sebbene, come notato, le ipotesi aprano a interessanti riflessioni sul legame tra sensazione di disgusto e paura e la somiglianza all'umano dei robot antropomorfi, l'apparato sperimentale atto a verificarle e validarle ad oggi non risulta ancora formalmente adeguato: i risultati riportati dagli studi che abbiamo analizzato, quando presenti, risultano comunque deboli nel validare l'ipotesi e l'apparato sembra più analizzare le impressioni a posteriori dell'esperienza di fronte al robot che l'effettiva reazione percettiva. Come vedremo (§5.2.2), queste difficoltà sono principalmente legate alle teorie riguardo alla mente e alla cognizione adottate, che in questo caso si manifestano nella considerazione della percezione come un meccanismo di raccolta dati al fine di computazioni cognitive successive. Anche nel caso della considerazione dell'aspetto evolutivo, questo non gioca un ruolo centrale nella strutturazione dell'esperienza di fronte al robot, quanto invece nella costruzione del meccanismo di raccolta dei dati percettivi tipico dell'essere umano: in sostanza, le teorie adottate concordano nel dipingere la percezione come un meccanismo tipico dell'essere umano modellato dall'evoluzione e a problematicamente applicato in ogni contesto, compreso quello che prevede l'incontro con una replica umana. A fianco alle critiche relative alle modalità di verifica delle indagini sperimentali,

quindi, vogliamo sottolineare che il contesto dinamico e situato in cui gli incontri avvengono, ancora una volta, è un fattore che in nessuno degli studi citati viene preso in considerazione.

Vedremo anche nel caso delle spiegazioni cognitive, infatti, che il modo in cui è più comune valutare la presenza dell'UVE negli incontri con i robot è attraverso uno schermo, tramite la somministrazione di immagini modificate (*morphed images*) e di questionari successivi al momento di osservazione. Il processo, dunque, come rilevato nel caso dell'attribuzione di intenzionalità (§4.3), da essere volto a scoprire i meccanismi percettivi alla base del fenomeno, si trasforma nella verifica delle modalità di giudizio dell'esperienza. A seguito del percorso che abbiamo compiuto nei capitoli precedenti, possiamo senz'altro notare, come fatto anche da altri (Wang et al. 2015), che questo non sembra il modo migliore per testare le ipotesi sull'UVE, per almeno due motivi: da una parte, gli esperimenti analizzati non sembrano aver raggiunto l'obiettivo di rintracciare legami causali tra le ipotesi di partenza e l'effetto inquietante; dall'altra, c'è da considerare che non tutte le repliche altamente somiglianti all'umano mostrano tratti estetici considerabili ripugnanti o che fanno pensare a malattie trasmissibili, o ancora che ricordano la morte, tanto da indebolire fortemente le posizioni che individuano direttamente nei caratteri somatici degli artefatti la causa regina del fenomeno.

In sostanza, quindi, possiamo passare ad analizzare l'insieme delle spiegazioni cognitive che sono state fornite per il fenomeno dell'UVE, dal momento che quelle percettive sembrano fare appello a una dimensione di categorizzazione e valorizzazione cognitiva dello stimolo che non consente loro di esaudire l'intento di spiegare il fenomeno autonomamente.

#### 4.4.5.2 Spiegazioni cognitive

L'insieme delle spiegazioni cognitive cerca di dare conto del fenomeno dell'UVE prendendo in considerazione «a broader and more general range of cognitive processing that occurs later» (MacDorman et al. 2009: 696) rispetto ai meccanismi squisitamente percettivi. Nello stato dell'arte attuale, come si sarà intuito, le teorie cognitive sono quelle a cui si è prestata più attenzione e che, all'interno del dominio, sembrano essere considerate le più promettenti per spiegare il fenomeno (Wang et al. 2015; Wang & Rochat 2017; Kätsyri et al. 2015; Zhang et al. 2020; Wiese & Weis 2020; Vaitonyte et al. 2023). Anche per questo insieme di spiegazioni, l'origine comune viene rintracciata nell'articolo di Mori (1970), le cui affermazioni vengono interpretate all'occorrenza a supporto delle ipotesi avanzate. In particolare, fanno parte dell'insieme delle spiegazioni cognitive la *violation of expectation hypothesis* (Brenton et al. 2005; Moore 2012; Mitchell et al. 2011; Saygin et al. 2012; Urgen et al. 2018; Seyama & Nagayama 2007), la *categorical uncertainty hypothesis* (Ramey 2006; Yamada et al. 2013; Cheetham et al. 2011, 2014, 2015; Mathur & Reichling 2016), la *mind perception hypothesis* (Gray et al. 2007; Gray & Wegner 2012; MacDorman 2024; Stein & Ohler 2017) e la

*dehumanization hypothesis* (Wang et al. 2015, Wang et al. 2020; Wang & Rochat 2017; Angelucci et al. 2016; Zlotowski et al. 2013, 2014).

- *Categorical Uncertainty Hypothesis* (CUH). Dalle *review* comparative realizzate sul tema delle spiegazioni dell'UVE (Wang et al. 2015; Kätsyri et al. 2015; Wiese & Weis 2020), l'ipotesi dell'incertezza categoriale sembra la più adottata per spiegare l'emersione di un sentimento di inadeguatezza e imbarazzo di fronte a figure altamente somiglianti all'umano (Cheetham et al. 2011, 2013; Burleigh & Schoener 2015; Mathur et al. 2020; Wiese & Weis 2020; Moore 2012; Ramey 2006). L'ipotesi, che affonda le sue radici nelle teorie psichiatriche di Jentsch (1906/1997), propone che il sentimento di disagio sia causato dalla impossibilità di decifrare la natura dell'entità antropomorfa, legata a un conflitto cognitivo che si genera in termini di convocazione concettuale per la categorizzazione di input percettivi discordanti. In altre parole, la CUH

[...] purports that the physical appearance of humanoid agents triggers a categorization-related cognitive conflict as to whether the agents represent human or nonhuman entities, and this conflict may result in negative emotional evaluations due to increased cognitive processing costs needed to resolve categorical ambiguity. (Wiese & Weis 2020: 1)

Dato l'alto grado di somiglianza presentato dalla creatura artificiale, secondo questa ipotesi il problema sorge perché il soggetto di fronte al robot è in difficoltà nel decidere se si trova di fronte a un artefatto o a un essere umano<sup>175</sup>. L'ipotesi è basata su una serie di studi provenienti dal dominio psicologico che individuano un confine categoriale nella percezione e identificazione delle immagini che raffigurano volti su un continuum che va da completamente artificiale a completamente umano e che l'azione di discriminazione tra entità umane e non umane è più dispendiosa in termini cognitivi quando ci si trova a disambiguare immagini nel mezzo o nei pressi della metà del continuum (Cheetham et al. 2011, 2013, 2014; Looser & Wheatley 2010; Yamada et al. 2013). L'incremento dell'incertezza categoriale è stato registrato coincidere con una valutazione negativa dello stimolo in entrata (Burleigh et al. 2013; Ferrey et al. 2015; Yamada et al. 2013), ma altri studi non sembrano supportare tale coincidenza (Cheetham et al. 2014; Looser & Wheatly 2010; MacDorman & Chattopadhyay 2016). Uno degli studi più importanti è quello realizzato da Moore (2012), il quale, adottando un modello bayesiano di percezione categoriale, ha ipotizzato che l'UVE sia un esempio di come la percezione possa essere distorta dai processi di categorizzazione, specialmente in presenza di stimoli che presentano indizi percettivi contrastanti riguardo alla loro appartenenza a una categoria. La tensione percettiva causata da questi segnali

---

<sup>175</sup> Una variazione di questa ipotesi prende il nome di *Perceptual Mismatch Hypothesis* (PMH), che afferma che «negative affinity associated with “uncanny” stimuli would be caused by an inconsistency between human-likeness levels of specific sensory signals contained in nonhuman images, such as grossly enlarged eyes displayed on an otherwise perfectly human-like face (MacDorman et al. 2009)» (Wiese & Weis 2020: 1). I presupposti che nutrono questa ipotesi sono gli stessi alla base dell'CUH, con cui condividono non solamente i principi teorici e le metodologie di indagine sperimentale, ma anche le criticità.



contrastanti si manifesta come disagio fisico o emotivo, come sensazioni di inquietudine o repulsione che scaturiscono dalla difficoltà di processazione dello stimolo in entrata nel meccanismo computazionale della cognizione.

The key, therefore, is the realization that, in the situation where there are multiple perceptual cues to category membership, there is the possibility that the multidimensional perceptual distortions induced at category boundaries could be misaligned. It is thus hypothesized that conflicting perceptual cues can give rise to differential distortion in the region of a category boundary, and that such distortion would be manifest as a form of perceptual ‘tension’. The idea is that such tension may be experienced as physical or emotional discomfort, e.g. feelings of eeriness or creepiness. (Moore 2012: 2)

È interessante sottolineare che, anche se l’articolo di Moore (2012) è uno degli studi più citati in materia di UVE, il suo contributo è relativo a una modellizzazione matematica della curva ipotizzata da Mori e il suo risultato più significativo consiste nel constatare, a seguito di una formalizzazione in termini formali e statistici, che «the model reveals that cue conflicts can be manifest as variations in the means and/or variances of their associated distributions or, more interestingly, from unequal levels of uncertainty associated with observing the different perceptual cues» (Moore 2012: 2). Ciò che lo studio fa, quindi, non è tanto validare l’ipotesi dell’UVE per come formulata da Mori, ma fornire una prova matematica del fatto che il grafico di Mori è compatibile con una formalizzazione del fenomeno in termini statistici che include l’ipotesi della CUH, soprattutto nella sua variante percettivamente motivata tramite *mismatch* (cfr. nota 173).

A parte le analisi matematiche, abbiamo più volte fatto riferimento a un diffuso metodo di verificare sperimentalmente l’ipotesi dell’UVE come metodologia più adottata, ovvero quella delle *morphed image*. La CUH è senza dubbio quella che più si è avvalsa di questo metodo di indagine, sia per confermare l’esistenza di una valle del perturbante, sia per comprendere i suoi meccanismi interni. L’osservazione di un’analisi in particolare può rendere chiaro l’impiego di questa metodologia di sperimentazione. Cheetham e colleghi (2011) hanno condotto uno studio in cui, attraverso la già nota metodologia delle *morphed images*, sono stati presentati stimoli provenienti da un continuum in cui un volto artificiale (“avatar”) viene progressivamente trasformato in un volto umano. In particolare, i ricercatori propongono che «the ability to discriminate between physically different stimuli might not be the same at all points along the (dimension of humanlikeness) DOH, with enhanced sensitivity for objects closest to the category boundary» (Cheetham et al. 2011: 2). In sostanza, secondo questa ipotesi, l’UV dovrebbe manifestarsi proprio quando si osservano immagini provenienti dalla parte centrale del continuum, dove c’è incertezza su quale categoria assegnare loro. Per verificare questa ipotesi, nel primo studio è stata selezionata casualmente un’immagine dal continuum e ai partecipanti è stato richiesto di identificarla rapidamente come appartenente alla categoria “avatar” o “umano”. Nel

secondo studio, ai partecipanti sono state presentate due immagini, con il compito di stabilire se appartenessero alla stessa categoria (“avatar” o “umano”), a categorie diverse (“avatar” e “umano”), oppure se fossero la stessa immagine. Arrivando ai risultati, questi non hanno pienamente confermato le ipotesi iniziali. Sebbene fosse emersa una maggiore incertezza nella valutazione di alcune immagini, infatti, queste non corrispondevano al confine categoriale: il grado più elevato di incertezza è stato osservato all’interno della categoria umana (Fig. 5.4), in seguito identificato con una percentuale di *morphing* tra il 60 e l’80% (Wiese & Weis 2020: 2). In altre parole, il confine tra la categoria “avatar” e la categoria “umano”, che dovrebbe rappresentare il punto di maggiore difficoltà nella discriminazione della categoria – e quindi il maggiore senso di incertezza – che provocherebbe la sensazione di inquietudine, è spostato rispetto alle performance registrate dai soggetti nell’esperimento. Questa evidenza non è isolata e anzi costituisce la maggioranza dei casi negli studi che hanno provato a testare la CUH. Altri esempi (Yamada et al. 2013; Wiese & Weis 2020; Seyama & Nagayama 2007; Cheetham et al. 2015; Mathur & Reichling 2016), infatti, rintracciano uno spostamento verso il versante dell’umano del punto di sforzo maggiore nella categorizzazione, disallineato rispetto a quello che dovrebbe corrispondere al confine tra le categorie concettuali. L’evidenza sperimentale, quindi, pone problemi di non semplice risoluzione ai sostenitori dell’ipotesi della CUH: qual è il motivo di questo spostamento? È possibile giustificare comunque l’UVE attraverso l’ipotesi dell’incertezza categoriale? Alcuni hanno suggerito che parte del problema potrebbe essere ricondotta alle modalità sperimentali con cui le ipotesi vengono verificate, ovvero le *morphed images*, che aggiungerebbero fattori ulteriori di confondimento e sposterebbero il confine categoriale verso l’umano (Kätsyri et al. 2015). Vedremo come, adottando un diverso concetto di percezione e di cognizione, sarà possibile evitare voli pindarici per rendere conto dell’esperienza di fronte al robot, soprattutto sostituendo le modalità di verifica sperimentale. Per ora vogliamo sottolineare che, come succederà anche per le altre teorie raccolte sotto le spiegazioni cognitive, la CUH non esplicita chiaramente il legame causale tra i meccanismi cognitivi ritenuti responsabili dei fenomeni alla base dell’UVE e l’UVE stesso. Nello specifico, per la CUH non viene reso chiaro dai sostenitori dell’ipotesi il motivo secondo cui un maggiore sforzo cognitivo nella discriminazione dello stimolo percettivo comporterebbe l’insorgenza di sentimenti negativi: in altri casi dello stesso tipo, in cui c’è bisogno di maggiore sforzo per riconoscere un qualcosa come tale (ad es. quando non capisco se una pianta è finta o vera) non sembra emergere alcuna sensazione negativa nei confronti dell’oggetto. Come si spiegherebbe, quindi, il rapporto tra l’incertezza categoriale e quello che avviene nell’UVE?

- *Violation of Expectation Hypothesis* (VEH). Quello della discordanza tra input percettivi, aspettative generate ed effetti creati dall’entità artificiale è considerato uno dei legami più

significativi per descrivere i meccanismi che genererebbero la sensazione di inquietudine, sullo stesso piano di plausibilità della CUH. Anche l'ipotesi della violazione delle aspettative affonda le sue origini nella formulazione originale del fenomeno. Infatti, Mori (1970/2005: 99-100), nell'esempio della mano protesica, aveva individuato l'emersione del fenomeno nell'incongruenza tra stimolo percettivo dato dalla mano, altamente somigliante a quella umana, e le informazioni tattili che potevano derivare dalla sua manipolazione. La VEH ipotizza che le repliche con un altissimo grado di *humanlikeness* suscitino l'effetto *uncanny* poiché creano delle aspettative nell'essere umano, grazie al fatto che sembrano umani, per poi però deluderle quando più accuratamente osservate o con il loro comportamento. A differenza di quanto accaduto con le spiegazioni percettive, la verifica di questa ipotesi è stata attuata in numerosi esperimenti, anche se con differenti teorie a supporto dell'interpretazione dei risultati. Ad esempio, Brenton e colleghi (2005) si sono concentrati sulla discrepanza tra gli stimoli percettivi prodotti da avatar, che provocherebbero l'UV dato il disallineamento tra elementi grafici e comportamentali: mentre l'aspetto sembrerebbe molto simile a quello umano, il comportamento generato dall'avatar non riuscirebbe a raggiungere lo stesso grado di accuratezza. Una parallela serie di studi, inoltre, ha ipotizzato che la violazione delle aspettative possa essere rintracciata in una ampia varietà di *mismatch*, come quelli fra l'apparenza e il movimento (Saygin et al. 2012), tra il volto e la voce del robot (Mitchell et al. 2011), o tra diverse caratteristiche del volto, come ad esempio sopracciglia artificiali e occhi umani (MacDorman et al. 2009; Seyama & Nagayama 2007). Tra queste, un'interessante linea di ricerca si è concentrata sull'applicazione del paradigma del *predictive coding* come teoria che possa dare conto della disattesa delle aspettative di fronte al robot e, quindi, dell'emersione dell'effetto *uncanny* (Saygin et al. 2012; Urgan et al. 2015, 2018). In particolare, Saygin e colleghi (2012) sono stati i primi a impiegare la tecnica di fMRI *repetition suppression*, che registra la risposta neurologica ridotta a uno stimolo ripetuto, per indagare il ruolo dell'*Action-Perception System* (APS), ovvero quel network cerebrale che si considera responsabile della percezione e dell'interpretazione delle azioni degli altri, nell'esperienza di fronte a robot con diversi gradi di somiglianza all'umano. In particolare, lo studio applica la tecnica di *neuroimaging* al fine di comprendere se l'APS si attivi in modo differente quando il soggetto osserva azioni compiute da un umano (aspetto biologico, movimento biologico), da un robot dalle caratteristiche meccanomorfe (aspetto meccanico, movimento meccanico) o da un androide (aspetto biologico, movimento meccanico). I risultati sono stati interpretati all'interno della teoria del *predictive coding* (Friston 2005, 2010; cfr. §5.2.2), suggerendo che il *mismatch* che si rintraccia nell'androide tra il suo aspetto e le sue capacità di movimento possa condurre ad un errore di previsione maggiore. Nel secondo gruppo di esperimenti, questi risultati sono stati rinforzati tramite l'uso

dell'elettroencefalografia (EEG) applicata a soggetti impegnati nella visione di immagini e video che ritraevano umani, robot meccanomorfi e androidi, i cui movimenti erano realizzati in accordo alle condizioni dell'esperimento precedente: focalizzandosi sul componente N400, solitamente adoperata in letteratura per testare l'ipotesi della predizione dell'errore (Urgen et al. 2015: 3), i risultati hanno confermato una maggiore attivazione quando il soggetto sta osservando azioni performati da androidi (aspetto biologico, movimento meccanico), supportando l'ipotesi che «uncanny valley might at least partially be due to violations of the brain's internal predictions about almost-but-not-quite-human robots» (Urgen et al. 2015: 1). L'esperimento è stato poi replicato, producendo gli stessi risultati (Urgen et al. 2018). All'interpretazione già fornita, però, nel secondo studio i ricercatori aggiungono una esplicita formulazione delle loro ipotesi in merito alla corrispondenza tra la predizione dell'errore e l'emersione dell'UVE:

[...] we suggest that the form (appearance) of the visually presented agent provided a context from which the subject inferred how the agent would move over time (e.g., a mechanical/robotic appearance would activate the semantic network that includes motion information associated with robotic appearances), the same way a preceding word group provides a context for the upcoming word in sentence (e.g. I take coffee with cream and \_\_\_\_ ("sugar" instead of "dog")), and activates the relevant semantic network associated with the word group. Thus, using EEG has allowed us to link the uncanny valley phenomenon to general cognitive processing using the well-established dependent measure N400. (Urgen et al. 2018: 185)

È importante notare che, anche se questa ipotesi risulta tra le più promettenti negli intenti e una delle più complesse in termini di fattori considerati, non riesce a spiegare perché i disallineamenti riscontrati debbano necessariamente causare il senso di inquietudine che caratterizza l'UVE: infatti, non sempre, come testimoniato da numerosi esperimenti (cfr. Margoni et al. 2024), la violazione delle aspettative comporta una sensazione spiacevole, quanto invece una ricalibrazione delle aspettative stesse e un processo di aggiustamento delle componenti predittive. Una possibile soluzione a questo ostacolo interpretativo potrebbe essere che tale sensazione sia il risultato di un particolare tipo di *mismatch*, che accade solamente nella situazione in cui le aspettative vengono violate di fronte a un robot umanoide che viene considerato alla pari di un essere umano:

While many nonbiological phenomena can violate our expectations, the eerie sensation associated with the uncanny valley may be peculiar to the violation of human-directed expectations, which are largely subconscious. If androids are more likely to fall into the uncanny valley than mechanical-looking robots, the reason may be that our brains are processing androids as human. (MacDorman & Ishiguro 2006: 301)

Tuttavia, la VEH così come è formulata allo stato dell'arte attuale non riesce a rendere conto di questa precisa circostanza, dal momento che negli articoli che cercano di spiegare l'UVE in questo modo non viene specificato alcun motivo che condurrebbe all'emersione dell'effetto di inquietudine davanti al robot androide e non in altre occasioni di *mismatch* registrato. Vedremo in

seguito come la base di questa ipotesi possa costituire un terreno fertile per la costruzione di un approccio enattivista al fenomeno, passando prima per un processo di riconfigurazione della teoria su cui si fonda.

- *Mind Perception Hypothesis* (MPH). Questa ipotesi cerca di dare conto del fenomeno a partire da un'analisi di cosa rende l'umano riconoscibile nei rapporti sociali. Gray e colleghi (2007) hanno individuato due principali dimensioni che permettono di distinguere gli esseri umani da quelli non umani: l'*agency*, intesa come l'abilità di pianificare e compiere azioni, e l'esperienza, concepita come l'abilità di percepire ed esperire situazioni e cose. In una ricerca successiva, Gray e Wegner (2012) hanno ipotizzato che sia nello specifico la seconda caratteristica, l'esperienza, ad essere evidentemente mancante negli esseri non umani, elemento che d'altro canto definisce sostanzialmente la natura umana. Sulla base di tale differenza, in questo articolo i ricercatori hanno proposto che le repliche umane non sono inquietanti poiché non abbastanza realistiche da essere scambiate per esseri umani, né perché ci sono elementi di disturbo che aumentano la difficoltà di decidere se siano umani o no, ma perché «they are so realistic that they trigger the attributions of human minds, particularly subjective experience, to nonhuman entities» (Wang et al. 2015: 397). In questo senso, per gli autori è precisamente l'attribuzione di esperienza umana ad artefatti artificiali come gli androidi che innesca il sentimento di inquietudine, dal momento che ciò viola la regola che associa l'esperienza umana a un agente vivo: in sostanza, quando il soggetto realizza che sta attribuendo una mente e un'esperienza a qualcosa che, costitutivamente, non può averne, allora il sentimento di disagio si manifesta come reazione alla stranezza dell'attribuzione. In Gray e Wegner (2012), gli autori hanno condotto tre studi per validare la MPH. Nel primo, hanno mostrato ai partecipanti le immagini di un robot umanoide o dall'aspetto meccanico e hanno chiesto loro di valutare su una scala da 1 a 5 alcune frasi che descrivevano i robot rappresentati in termini di "stranezza" (*uncanniness*) e della capacità di percepire e sentire (ad es., "This robot has the capacity to feel pain") o per la capacità di agire e pianificare ("This robot has the capacity to plan action"): i risultati in questo caso hanno mostrato valutazioni più alte di *uncanniness* per i robot più umanoidi. Nel secondo studio, hanno cercato di comprendere se una forte somiglianza all'umano fosse necessaria per l'emersione dell'UVE, anche qui attraverso un questionario in cui il partecipante valutava le possibilità di esperienza di alcuni computer con diverse capacità descritte dagli autori: le ipotesi sono validate dal momento che l'UVE viene riscontrata anche di fronte ad artefatti non umanoidi, rendendo la somiglianza all'umano un fattore non strettamente legato a livello causale all'UVE. Infine, nell'ultimo studio i ricercatori hanno ipotizzato che anche la percezione di una mancanza di esperienza da parte del robot potesse suscitare l'UVE e, attraverso la somministrazione di un altro questionario, anche questa ipotesi è stata interpretata

come valida. In questo, la MPH prende in prestito l'idea della disattesa delle aspettative così come proposta dalla VEH<sup>176</sup> e propone che le cause dell'UVE risiedano i) nella attribuzione di esperienza, caratteristica costitutivamente umana, al robot e ii) dalla disattesa delle aspettative da parte della replica, che viene considerata avere un'esperienza ma poi viene scoperta non esserne costitutivamente in grado.

Ora, sebbene anche questa ipotesi sia decisamente interessante e possa costituire un notevole punto di partenza per riflettere sul fenomeno dell'UVE dal punto di vista interazionale, le evidenze non sono molto chiare. Anche qui, infatti, il legame causale sembra riportare non pochi problemi: secondo i ricercatori, sia l'attribuzione di esperienza al robot, sia la disattesa delle aspettative da parte del robot circa la sua capacità effettiva di avere un'esperienza sono interpretabili come cause dell'UVE. Tuttavia, secondo noi non è possibile annoverare entrambe le ipotesi causali all'interno di un unico approccio al fenomeno basato sull'esperienza, a meno di fornire una adeguata giustificazione che lega i risultati dei fenomeni e spiega l'esito comune: se l'essere umano prova sensazioni di disagio nel rendersi conto che sta attribuendo esperienza a qualcosa che sa non poterne avere, allora quando scopre di "essere in torto" dovrebbe sentirsi sollevato, o quanto meno ricalibrare il proprio insieme di credenze affinché le predizioni si rivelino più accurate; allo stesso tempo, se l'UVE dipende dalla disattesa delle aspettative circa la capacità di esperire, allora il momento di attribuzione di esperienza non può avere il portato emotivo tipico dell'UVE. Anzi, numerose evidenze che abbiamo annoverato nel capitolo III (§3.4.1) legano al momento di attribuzione di esperienza, di mente o di altri tratti ritenuti essere caratteristici dell'umano, l'attivazione dell'antropomorfismo, fattore che notoriamente permette – e non ostacola – l'interazione con artefatti artificiali.

- *Dehumanization Hypothesis* (DH). L'ipotesi della deumanizzazione può essere considerata la più recente nel novero delle teorie cognitive che vogliono spiegare l'UVE, ma soprattutto negli ultimi anni sta guadagnando terreno come plausibile elemento esplicativo. Il nucleo centrale dell'ipotesi si basa sia sulla disattesa delle aspettative, sia sull'attribuzione di mente, e suggerisce che l'UVE sia causato da un processo percettivo in cui il robot arriva ad essere visto come non completamente umano: dapprima, la naturale tendenza all'antropomorfizzazione innescerebbe una attribuzione di mente al robot, a cui seguirebbe una *sottrazione* (Wang et al. 2015, 2020; Wang & Rochat 2017) della stessa quando l'umano realizza che quello a cui viene attribuita un'esperienza e una mente non è umano. È qui, dunque, che si verifica il processo di deumanizzazione del robot, che prima di tutto deve acquisire uno statuto "umano" che subito dopo gli viene negato a causa della

---

<sup>176</sup> Vedremo nel paragrafo successivo come la MPH e la VEH non condividano esattamente lo stesso tipo di teoria della percezione.

realizzazione del soggetto. Seguendo gli studi sul fenomeno in ambito antropologico e sociologico (Haslam 2006), il processo di deumanizzazione si configura così come il motivo scatenante dell'UVE, un richiamo corporeo ed affettivo che segnala la presenza di un errore e si sviluppa in corrispondenza di una elaborazione cognitiva di alto livello di uno stimolo percettivo ingannatore computato come non si sarebbe dovuto. Secondo i ricercatori, le tempistiche in cui questo fenomeno si realizza sono velocissime, tra i 100 e il 1000 millisecondi dalla presentazione dello stimolo: i ricercatori ipotizzano che il fenomeno si realizza intorno ai 400 ms dalla visione del robot, momento in cui il soggetto distingue il volto dell'androide da quello di un essere umano.

In particular, the rapid face detection coupled with a natural tendency for anthropomorphism predicts an initial overattribution of minds to android faces, followed by a subtraction of minds from them when humans discriminate android from human faces beyond 400ms after stimulus onset. (Wang et al. 2020: 1071)

L'ipotesi si sviluppa attraverso tre elementi chiave: i) si focalizza sul livello algoritmico di analisi, con l'obiettivo di specificare i processi cognitivi che conducono alla presentazione dell'UV; ii) si fonda sul meccanismo di percezione e attribuzione di mente come elemento fondamentale nella costituzione dell'UVE; iii) si propone come declinazione sia della VEH, in quanto analizza i modi in cui le aspettative vengono violate, sia della MPH, poiché parte dagli stessi presupposti teorici ed epistemologici radicati nell'antropomorfismo<sup>177</sup>. In sostanza, la DH fornisce un livello di analisi più dettagliato dei modi in cui le aspettative e l'attribuzione di mente vengono disattese, soprattutto in riferimento alla *face animacy perception*.

In particolare, lo studio di Wang et al. (2020) fornisce un'interessante prova del ruolo che la percezione dell'*animacy* – ovvero la capacità di oggetti e entità di mostrarsi “vivi” – nei volti ha nella diversa considerazione di umani e non umani. Nel primo esperimento, i ricercatori hanno esaminato l'effetto del tempo di esposizione a volti umani, androidi e robot meccanici. I partecipanti hanno osservato immagini statiche di 57 volti appartenenti a queste diverse categorie

---

<sup>177</sup> Una precisazione è, tuttavia, dovuta. Nell'articolo di Wang e colleghi (2020: 1071), gli autori chiariscono che la DH e la MPH condividano alcuni assunti circa il processo generale, ma divergano nella considerazione della struttura del fenomeno: « Although both the dehumanization and the mind perception hypotheses argue for the crucial role played by mind perception in eliciting the uncanny valley, the two hypotheses differ profoundly in their underlying assumptions, which lead to opposing predictions regarding how perceived face animacy would change as a function of exposure time. In particular, the dehumanization hypothesis assumes that people initially attribute high animacy to an android face, predicting a decrease in animacy over time following 400 ms after stimulus onset. In contrast, the mind perception hypothesis is based on the default assumption that an android face is inanimate; only with increasing exposure to the face do people attribute a mind to it, predicting an increase in animacy over time (Wang et al., 2015). Therefore, according to the dehumanization hypothesis, we predicted that perceived animacy should decrease in android faces. In contrast, we predicted that perceived animacy should remain stable in human and mechanical-looking robot faces». Conveniamo che le concezioni siano differenti, ma allo stesso tempo Wang e colleghi sembrano estremizzare la posizione della MPH riguardo al fatto che, se non a seguito di prolungate esposizioni, il volto del robot è percepito come inanimato. In realtà, negli studi sulla MPH non si fa riferimento a una dimensione temporale precedente alla comprensione dello stimolo (cfr. Gray et al. 2009; Gray & Wegner 2012): in sostanza, la spiegazione non sembra del tutto giustificata dai riferimenti citati, fattore che ci permette di tenere insieme i due approcci sotto il fenomeno dell'antropomorfismo, a cui sia Wang e colleghi (2015, 2020; Wang & Rochat 2017), sia Gray e colleghi (2009; Gray & Wegner 2012) si appellano.

per 100ms, 500ms e 1000ms. Dopo ogni esposizione, i soggetti dell'esperimento dovevano valutare su una scala da 1 a 6 quanto il volto apparisse "vivo" (indice di animazione percepita). I risultati hanno mostrato che volti di umani e di robot meccanici hanno mantenuto valutazioni costanti di *animacy* a prescindere dal tempo di esposizione, mentre i volti androidi hanno mostrato una significativa diminuzione della percezione di *animacy* tra 100ms e 500ms, riduzione stabilizzata tra 500ms e 1000ms. Questi risultati sono letti a supporto dell'ipotesi della deumanizzazione: inizialmente, gli androidi vengono percepiti come "vivi", ma se considerati in un'esposizione più lunga vengono "deumanizzati". Nel secondo studio, poi, i soggetti sono stati esposti a volti che erano stati sottoposti a una manipolazione dei tratti, in particolare relativi a informazioni di *Low Spatial Frequency* (LSF) – tratti estetici considerati fornire dati grezzi a supporto della processazione cognitiva dei volti – e di *High Spatial Frequency* (HSF) – che invece individuano dettagli e particolari di un volto. L'ipotesi è supportata da una serie di studi (Looser & Wheatley 2010; Peters et al. 2018) che rilevano un modello a due passaggi nella processazione del volto, in cui queste due frequenze giocano ruoli diversi nell'attribuzione dell'*animacy*. Dunque, i ricercatori hanno mostrato ai soggetti immagini con frequenze spaziali modificate: hanno conservato alternativamente la LSF e la HSF dei volti degli androidi durante tempi di esposizioni prolungati per verificarne l'influenza sulle percezioni di *animacy* e *uncanniness*, con l'obiettivo di vedere se l'eliminazione di alcune informazioni percettive riducesse l'UVE. I risultati, raccolti anche in questo caso tramite questionari, hanno confermato che eliminare le HSF, mostrando quindi un volto più stilizzato e con un minor grado di somiglianza all'umano, comporta una riduzione del senso di inquietudine, ma anche una diminuzione della *animacy* percepita. Oltre alla validazione delle proprie ipotesi, Wang e colleghi (2020) hanno dimostrato che i tempi di esposizione ricoprono un ruolo cruciale nell'emersione dell'UVE e della percezione del robot come animato (cfr. Zlotowski et al. 2015).

L'ipotesi sembra essere promettente sia dal punto di vista filosofico (Angelucci et al. 2014), sia sperimentale, fattore che contribuisce alla attuale popolarità della spiegazione, anche se alcuni aspetti restano ancora da chiarire. Gli stessi autori dello studio citato dichiarano alcune difficoltà nella misurazione degli stretti tempi di esposizione e ammettono la opinabilità di verifica degli studi su cui la DH è basata (Wang et al. 2020: 1082). A questo, possiamo aggiungere almeno due ulteriori aspetti che ci paiono problematici. In primo luogo, anche in questo caso torna a essere centrale la struttura del legame causale tra i fenomeni studiati e l'emersione di una sensazione di inquietudine: il motivo per cui la deumanizzazione del robot, a seguito dell'attribuzione di una mente, causi l'UVE non è esplicitato, quanto invece lasciato sospeso: l'UVE è una diretta conseguenza di questo meccanismo? Avviene solamente legato a contesti in cui è preceduto da un



corrispettivo di antropomorfizzazione, o si realizza anche indipendentemente da questo? Se fosse il caso della prima ipotesi, la DH non potrebbe poggiarsi sulla MPH, come invece succede, dal momento che quest'ultima concepisce il momento di attribuzione della mente al robot come la causa diretta dell'UVE e non il suo presupposto. In secondo luogo, il fatto di porre l'attenzione solamente sul volto comporta una significativa perdita di contesto, che ovviamente non fa altro che rinforzare l'idea che l'interazione situata non abbia rilevanza negli impianti sperimentali. È assodato che il volto ricopra un ruolo fondamentale nella percezione reciproca e nella comunicazione interpersonale (Gallagher 2008b; Stawarska 2010; Zlatev et al. 2008 (Eds.); Frith 2009), ma questo non significa che sia l'unica parte dell'agente che conti: il processo di deumanizzazione potrebbe trovare nuova luce con gli studi che considerano l'integrità del corpo dell'agente in interazione e, soprattutto, che conferiscono un'importanza sostanziale al modo in cui questo si comporta nel mondo e con gli altri agenti.

D'altra parte, questo tipo di studi ha contribuito a introdurre nel dibattito sull'UVE un fattore che fino a pochi anni fa era totalmente ignorato: l'UVE è un fenomeno multidimensionale e processuale, che si sviluppa nel tempo e che convoca nella sua realizzazione una pluralità di processi cognitivi e percettivi, che si conseguono e che si influenzano a vicenda. Questa concezione del fenomeno permette di allargare l'indagine, prendendo in considerazione non solamente un meccanismo alla volta allo scopo di rintracciare il nodo cruciale che scatena il fenomeno nella sua totalità, ma pensando all'UVE come un fenomeno dallo sviluppo dinamico che si dispiega *nell'incontro* degli agenti coinvolti. Nel prossimo paragrafo, vedremo come questo presupposto sia la chiave per aprire a una visione dell'UVE come il risultato *in itinere* dell'interazione, invece che la fine di un incontro andato male.

In sostanza, anche le spiegazioni cognitive mostrano una serie di criticità che non consentono ad alcuna di loro, almeno individualmente, di rendere conto del fenomeno dell'UVE nella sua complessità. I problemi più consistenti sono rintracciabili: i) nella debolezza del legame causale tra i fenomeni considerati essere la causa dell'UVE e l'emersione del senso di inquietudine; ii) nella metodologia sperimentale scelta per verificare l'ipotesi nella maggior parte degli esperimenti, che nonostante vogliano indagare un effetto che emerge di fronte a un robot, prevede principalmente l'osservazione di immagini statiche, fotografie o avatar (eccezion fatta per gli studi a supporto della VEH, in cui vengono utilizzati dei video come input percettivi); iii) nell'inconsistenza dei risultati, che si mostrano spesso contraddittori anche all'interno della stessa ipotesi difesa (come nel caso della CUH) e che rendono la situazione fumosa e di difficile interpretazione. Dall'altra parte, possiamo raccogliere una serie di elementi interessanti, che provengono dalle tesi più innovative sulle cause

dell'UVE e che possono essere considerate date le premesse che abbiamo posto circa la volontà di costruire un approccio enattivista a questo *breakdown* dell'interazione tra umani e robot: a) in primo luogo, spiegazioni come la MPH e la DH mettono in primo piano la natura processuale e complessa del fenomeno dell'UV, chiamando in causa fattori che esulano dalla ristretta visione appartenente alla percezione categoriale o al confinamento del fenomeno nei parametri di una risposta evolutiva a stimoli percettivi particolari; b) dall'altra, tesi come la VEH mettono in luce il fatto che la complessità del fenomeno possa essere determinata anche dal modo in cui il robot si *comporta* nell'ambiente, e non solo da come appare.

Ciò che ci aspetta, allora, è cercare di trovare una soluzione comune a questi punti critici ancora tutti da chiarire: partiremo da un'analisi della radice comune a tutti questi interrogativi, a cui faremo seguire un tentativo di spiegazione che tenti di tenere insieme i) la particolarità dell'UVE, un *breakdown* che sembra emergere, almeno nella forma più esplicita e potente, solamente nella HRI con robot umanoidi o androidi; ii) la struttura complessa del fenomeno, che sembra evolversi durante l'interazione e non presentarsi come un effetto a monte di questa; iii) il ruolo costitutivo dell'interazione, tralasciato dalla totalità degli studi che abbiamo analizzato ma che abbiamo visto costituire le fondamenta del modo in cui il robot è considerato un agente intenzionale dall'essere umano che si trova ad interagire con esso.

#### 4.4.6 *Sull'impostazione teorica e metodologica degli studi sull'UVE*

È possibile notare che le diverse spiegazioni dell'UVE, nel tentativo di fornire ragione della sensazione di inquietudine che emerge di fronte a robot con un alto grado di somiglianza all'umano, fanno tutte riferimento a un processo che va dal percettivo al cognitivo, il quale consta di diverse fasi. Ciascuna spiegazione fornita fin qui si poggia, infatti, su un approccio alla cognizione di tipo funzionalista e consiste in un processo di raccolta delle informazioni, le quali vengono elaborate cognitivamente da meccanismi computazionali dediti a questo compito specifico e rintracciabili in un modulo nella struttura funzionale del cervello. In un modo o nell'altro, sia le percezioni cognitive sia quelle percettive rintracciano le cause dell'UVE in qualche malfunzionamento del processo di computazione dello stimolo percettivo, che questo si collochi alla fine, nel mezzo o a monte di questo processo: l'adozione di diverse teorie della cognizione, quindi, trova il proprio minimo comun denominatore in un approccio che vede la percezione come l'attività di raccolta passiva degli stimoli esterni (bottom-up) e la cognizione come l'applicazione di meccanismi computazionali in grado di interpretare lo stimolo (top-down).

Nello specifico, a eccezione della VEH che analizzeremo in seguito, tutte le altre teorie si basano su un modello input-output della cognizione, fondato su un funzionalismo causale modularista (Fodor

1983; Putnam 1967; Block 1995; Bechtel 2003) che parte dalla raccolta di informazioni percettive in entrata e passa per altri moduli cognitivi, i quali elaborano i dati grezzi e producono una rappresentazione mentale come risultato del processo percettivo, il quale può essere anche inviato ad altri moduli che elaboreranno ulteriormente la rappresentazione a seconda della funzione a cui sono preposti. A seconda dell'ipotesi, l'UVE viene spiegata attraverso il funzionamento o il malfunzionamento di alcuni di questi moduli, responsabili in modi diversi dell'acquisizione di significato di cui si dota l'oggetto intenzionato.

In questo quadro, le spiegazioni percettive (PAH, MSH, EAH) ritengono che l'UVE dipenda dal modo in cui la rappresentazione mentale del robot, risultante dall'elaborazione dello stimolo percettivo in entrata, acquista un contenuto all'interno della struttura funzionale della mente dell'individuo. Questa elaborazione conduce a un'associazione che lega il riconoscimento della figura robotica a reazioni fisiologiche di disgusto, di percezione di morte o di incompatibilità interspecifica. Passando alle spiegazioni cognitive, d'altra parte, la MPH non individua problemi nella rappresentazione percettiva, ma nel fatto che tale rappresentazione attivi il modulo funzionale legato alla teoria della mente e che questa attivazione sia in conflitto con la conoscenza pregressa circa l'associazione tra attribuzione di esperienze e di agentività ed esseri viventi e il fatto che questo avviene nei confronti di un robot. La CUH, invece, individua il problema nel processamento dell'informazione all'interno del modulo percettivo<sup>178</sup>, ossia nel passaggio dallo stimolo bruto alla sua percezione categoriale: qui, il nodo cruciale viene individuato nel momento in cui il modulo cognitivo applica gli algoritmi che permettono, solitamente, di discriminare tra diversi stimoli percettivi in entrata, i quali dovrebbero associare il percelto a un contenuto rappresentazionale identificato con una categoria precisa e generare, quindi, una rappresentazione coerente dell'oggetto. Secondo la CUH, la somiglianza all'umano da parte del robot fa intoppiare il meccanismo e mette il processo cognitivo modulare in difficoltà: come l'asino di Buridano<sup>179</sup>, il processamento rimane invischiato nell'operazione di decidere per una delle due alternative che sembrano plausibili, nella scelta ambigua tra le categorie concettuali di umano e non umano. La DH, invece, individua il

---

<sup>178</sup> Il modello di riferimento di questa teoria, anche se non esplicitato, è chiaramente quello di Marr (1982; Marr e Poggio 1979), che costituisce la teoria della percezione più affermata all'interno delle scienze cognitive classiche e contemporanee (Marraffa e Paternoster 2012). In sintesi, secondo questo approccio, la percezione funziona attraverso una serie di stadi in successione che trasformano l'informazione sensoriale grezza in rappresentazioni cognitive. Il processo inizia con l'estrazione delle caratteristiche di base, come bordi e contorni, nel livello computazionale, dove viene definito lo scopo della percezione: comprendere l'input e riconoscerlo. Successivamente, al livello algoritmico, questi dati vengono elaborati per costruire rappresentazioni tridimensionali dell'oggetto, utilizzando algoritmi che integrano elementi come la profondità. Infine, al livello dell'implementazione, tali operazioni sono realizzate a livello neuronale, coinvolgendo strutture cerebrali come la corteccia visiva primaria e le aree associative, responsabili dell'integrazione delle informazioni sensoriali in una percezione visiva coerente e organizzata. Questo modello modulare e gerarchico della percezione è stato fondamentale nelle scienze cognitive classiche, fornendo un quadro strutturato che ha permesso di leggere i processi mentali complessi attraverso una lente funzionalista e computazionalista.

<sup>179</sup> Devo questa similitudine a Luigi Lobaccaro (comunicazione privata), che ha saputo riassumere in questa immagine la maggior parte delle critiche alle teorie della percezione categoriale elaborate in letteratura.

problema in una mancata integrazione tra le rappresentazioni cognitive provenienti da diversi moduli funzionali. In particolare, mentre il modulo percettivo costruisce una rappresentazione del robot basata sulle sue caratteristiche fisiche, il modulo responsabile della ToM genera aspettative legate a comportamenti tipici di agenti umani. La difficoltà di coordinare queste rappresentazioni disgiunte – una che dipinge il robot come umanoide e l'altra che non riconosce in esso i tratti agentivi e intenzionali propri degli esseri umani – porta a una rottura nell'elaborazione cognitiva, che genera il processo di deumanizzazione.

Abbiamo già visto come, in realtà, nessuna di queste opzioni fornisca soluzioni ed evidenze soddisfacenti al problema dell'UVE. A questo punto, è fondamentale mostrare come questi approcci conducano a due ordini di problemi teorici circa l'individuazione e la concettualizzazione dell'UVE:

- i) In primo luogo, in tutti i casi il problema dell'UVE viene considerato scaturire dal modo in cui la cognizione computa gli stimoli percettivi raccolti dal soggetto, non considerando la possibilità che la chiave per comprendere il fenomeno possa essere rintracciata non solo in ciò che pertiene alle capacità cognitive e interpretative dell'individuo. Il focus sui meccanismi cognitivi, infatti, sposta l'attenzione tutta sul processo di interpretazione del dato, mentre la percezione viene considerata solamente come un'azione passiva di ricezione dello stimolo. Questo è chiaro sia per la PAH, in cui l'effetto risultante è la repulsione, sia per la MSH, che vede l'UVE generata dal senso di morte percepita e ricordata, sia per la EAH, in cui il disgusto è provocato dal riconoscimento di un consimile dalle inquietanti qualità estetiche. In tutti questi casi, anche se i processi possono avvenire a livello sub-personale senza arrivare allo stato di coscienza del soggetto, la percezione passa per una elaborazione ulteriore che permette l'intelligibilità dell'oggetto percepito. A maggior ragione, quando si considerano le spiegazioni cognitive questo aspetto viene reso chiaro: gli stimoli percettivi vengono riportati a meccanismi di interpretazione che rendono possibile il loro riconoscimento o la loro categorizzazione. Questa concezione è riscontrabile in tutti gli approcci all'UVE, ma risulta particolarmente chiara negli studi a supporto della CUH, le cui ipotesi riguardano esattamente il processo di categorizzazione degli stimoli percettivi. In particolare, nello studio di Cheetham et al. (2011) abbiamo visto che i ricercatori tentano di rendere conto dell'UVE a partire dall'assunto che le immagini mostrate nel continuum delle *morphed images* appartengano inequivocabilmente alla categoria dell'umano o a quella dell'avatar, fatta eccezione per quelle posizionate a metà tra gli estremi che dovrebbero suscitare l'UVE. In questo caso, l'effetto perturbante dovrebbe scaturire esattamente dall'eccezione alla regola della percezione basata sulla categorizzazione e sul riconoscimento. Come dimostrato dagli esperimenti, però, il

momento di maggiore tensione cognitiva che dovrebbe comportare l'UVE viene rintracciato non nel punto in cui era stato previsto ci dovesse essere il dubbio fra le categorie concettuali, elemento che pone in seria difficoltà i sostenitori dell'ipotesi dell'incertezza categoriale e, di conseguenza, le teorie che a questa si appoggiano per spiegare l'UVE. La modellizzazione della capacità percettiva come un'attività computazionale può essere sicuramente utile per realizzare intelligenze artificiali con abilità di riconoscimento dell'ambiente circostante, ma non sembra essere adeguata a rendere conto della moltitudine eterogenea di fattori implicati nel processo percettivo dell'essere umano, né a livello teorico, né nella specifica circostanza dell'UVE. Trascurare questo fattore conduce, da una parte, a produrre interpretazioni fantasiose per giustificare il disallineamento dei risultati sperimentali e, dall'altra, a ridurre il fenomeno studiato ad una errata applicazione dell'algoritmo della percezione.

- ii) In secondo luogo, considerare la percezione come meccanismo di raccolta degli stimoli atti a innescare il processamento cognitivo struttura l'analisi dell'UVE mettendo in primo piano il modo in cui il robot si presenta al soggetto in termini di sembianze. Infatti, tutte le ipotesi formulate per rintracciare e testare le cause dell'UVE attribuiscono alla dimensione della *humanlikeness*, ovvero della somiglianza dell'artefatto robotico all'umano, il ruolo di fattore scatenante del fenomeno. Le spiegazioni percettive partono proprio dalla pertinentizzazione dei tratti estetici del robot per giustificare l'emersione della sensazione di disagio: la PAH e la MSH rintracciano la causa nella necessità imposta dall'evoluzione di preservare la propria integrità, rispettivamente fisica e mentale, e l'*uncanniness* si manifesta come risposta affettiva che avverte l'individuo di un rischio proveniente dall'entità osservata; nella EAH, poi, lo specifico tipo di *humanlikeness* mostrata dai robot è esattamente ciò che scatena la reazione di disgusto o paura, ipotizzata come la risposta fisiologica a incompatibilità interspecie o specie-specifiche. Per le spiegazioni cognitive, il grado di somiglianza all'umano è altrettanto indispensabile per l'emersione dell'UVE. Salvo casi sporadici (Sierra Rativa et al. 2022, 2020; Schwind et al. 2018; Yamamoto et al. 2009), l'UVE è attribuita a un rapporto tra la *humanlikeness* e vari meccanismi cognitivi atti ad elaborarne il significato: per la CUH, l'UV si dovrebbe registrare in corrispondenza di un compito cognitivo dispendioso innescato dalla difficile elaborazione degli input percettivi provenienti dal volto del robot; la MPH considera la *humanlikeness* l'elemento scatenante del fenomeno di percezione di mente; per la DH, l'estetica del robot è ciò che innesci l'attribuzione di umanità e allo stesso tempo la delude. In realtà, l'idea comune riguardo la centralità della dimensione della somiglianza proviene direttamente dalla

formulazione originale di Mori, che individuava proprio nel crescere della *humanlikeness* il motivo scatenante dell'effetto *uncanny*. Tuttavia, quella di Mori si configurava come un'ipotesi speculativa, data l'assenza di prove scientifiche a supporto: il processo di messa alla prova delle ipotesi nei contesti sperimentali ha messo in luce, in molti casi, che la *humanlikeness* sia un elemento cruciale nell'emersione dell'UVE, ma anche che non sia l'unico. Altri fattori, come ad esempio la processualità della cognizione nell'incontro con il robot umanoide evidenziato nella DH, sono stati indicati come ugualmente fondamentali da considerare nel novero delle cause all'UVE. In questo contesto, il metodo sperimentale adottato nella maggior parte delle indagini sull'UV – che consiste nella somministrazione di immagini statiche della figura del robot o anche solamente del suo volto – non riesce a rendere conto della multidimensionalità e della pluralità dei fattori che generano l'UVE: utilizzare immagini statiche negli esperimenti per verificare la presenza e la conformazione del fenomeno anestetizza una serie di fattori legati al movimento e alla processualità dell'interazione dell'agente con il proprio ambiente, dal momento che consistono esclusivamente nell'osservazione di rappresentazioni iconiche dell'agente, spesso nemmeno a figura intera. Abbiamo spesso notato, infatti, che la maggior parte degli studi adotta un approccio che identifica nel volto la zona di interesse, per giunta congelato in una immagine artificiale o in una foto che viene sottoposta all'osservazione dei partecipanti all'esperimento, lasciando da parte tutto il resto: il contesto, l'esperienza immediata dell'interazione, la componente corporea e motoria, le influenze personali intime come la disposizione momentanea, il condizionamento affettivo, la predisposizione maturata nei confronti di macchine antropomorfe tramite esperienze precedenti non solo in prima persona, ma date anche dalle radici sociali e culturali dell'individuo.

Riflettendo su questi punti, ci si può chiedere se la difficoltà di individuare una spiegazione unitaria dell'UVE derivi esclusivamente da questioni metodologiche e di interpretazione dei risultati, o se i problemi che abbiamo riscontrato possano indurci a mettere in discussione l'intero impianto epistemologico chiamato a sostegno di queste ipotesi.

Alcune delle teorie cognitive che abbiamo annoverato a supporto della VEH sembrano andare proprio verso quest'ultima direzione (Saygin et al. 2012; Urgan et al. 2015, 2018). Infatti, abbiamo già accennato al fatto che la VEH trova la sua declinazione più interessante nel richiamo alle teorie cognitive del *predictive coding* (Howhy 2013; Clark 2013, 2016), le quali, pur condividendo un'impostazione funzionalista, ipotizzano un'organizzazione distribuita dei network neurali, abbandonando l'idea di una struttura modulare dell'architettura cognitiva e di una suddivisione in

unità computazionali dedicate a compiti cognitivi specifici. Non più rigidamente compartimentate in moduli, dunque, le reti neurali emergono come sistemi dinamici e adattativi, in cui l'elaborazione delle informazioni si sviluppa attraverso interazioni complesse e continue tra i vari nodi. Secondo questa concezione, la cognizione non è il risultato di processi isolati, che avvengono in stadi compartimentati e specifici del processo cognitivo, quanto invece il frutto di una sinergia tra diversi livelli di rappresentazione e di risposta alle stimolazioni ambientali. In questo contesto, il *predictive coding* offre un quadro teorico in grado di spiegare come il cervello modelli e anticipi le esperienze sensoriali, integrando le informazioni passate con quelle presenti per ottimizzare l'interazione con il mondo esterno. Negli ultimi anni, le indagini in merito alla modalità predittiva cerebrale si sono moltiplicate, fino a costituire uno dei paradigmi più diffusi nelle ricerche neurocognitive sulle funzioni del cervello e sui meccanismi che questo realizza nel processo di comprensione dell'esperienza umana (Frith & Frith 2006; Millidge et al. 2021; Bubic et al. 2010; Kirchhoff & Kiverstein 2019, 2020; Seth 2014; Friston 2005; Gallagher & Allen 2018). Questo ha fatto sì che le varianti delle teorie predittive si moltiplicassero e che venissero sviluppate in accordo a diversi *framework* teorici<sup>180</sup>.

Ciò che segna la maggiore differenza tra gli approcci funzionalisti classici e le teorie del *predictive coding* riguarda il compito pensato essere alla base del meccanismo cognitivo: se nel modello computazionalista, ad esempio quello di Marr (1982), lo scopo della percezione (nello specifico visiva) è quello di *riconoscere la forma geometrica* di un oggetto, nel *predictive coding* lo scopo primario della cognizione e dei meccanismi percettivi al suo servizio è quello di *minimizzare continuamente l'errore di predizione* (Friston 2005): il cervello, quindi, non si limita a elaborare lo stimolo percettivo in entrata, ma anticipa attivamente gli input sulla base di modelli interni, detti *prior*, cercando di ridurre la possibilità della discrepanza tra ciò che è previsto e ciò che viene effettivamente percepito (Clark 2013, 2016).

In particolare, il *predictive coding* adottato negli studi a supporto della VEH (Friston 2005, 2010) costituisce una variante computazionalista della teoria predittiva, in cui il cervello è visto come un sistema gerarchico che genera continuamente predizioni sugli input sensoriali attesi. Ogni livello della gerarchia cerebrale superiore produce una rappresentazione interna del mondo, basata su alcuni modelli predittivi. Queste predizioni vengono inviate ai livelli inferiori del sistema neurale, dove si confrontano con l'input sensoriale reale: quando l'informazione sensoriale percepita non corrisponde

---

<sup>180</sup> Per una panoramica sulla costellazione delle teorie relative alle modalità predittive del cervello, si veda Millidge et al. (2021); Seth (2014); Bubic et al. (2010). In seguito (§5.3.1), prenderemo in considerazione la declinazione enattivista del modello predittivo, affinché sia chiaro come non sia solamente il cervello a operare una previsione circa ciò che l'agente si deve aspettare, ma tutto il corpo attraverso lo scambio continuo di interazioni con l'ambiente in cui è inserito (cfr. Gallagher & Allen 2018; Kirchhoff & Kiverstein 2019; Allen & Tsakiris 2019; Bruinenberg et al. 2018) e il contributo dell'esperienza incarnata in un contesto sociale, culturale e narrativo (Gallagher et al. 2022; Hutto & Hipolito 2024).

alla predizione, si genera un errore di predizione, che rappresenta la discrepanza tra le aspettative e l'input effettivo. L'obiettivo principale di questo processo è minimizzare questo errore, inviando segnali di *mismatch* dai livelli inferiori a quelli superiori, consentendo così un aggiornamento continuo delle predizioni. Questo processo di minimizzazione dell'errore di predizione avviene attraverso un'inferenza *bayesiana*, nella quale le rappresentazioni interne vengono aggiornate sulla base della probabilità che una data predizione sia corretta. Se l'errore di predizione persiste o diventa significativo, il cervello modifica i propri modelli predittivi, migliorando la corrispondenza tra predizioni e percezioni future.

In sostanza, il processo può essere schematizzato seguendo il modello: input sensoriale (*bottom-up*) → confronto con la predizione (*top-down*) → generazione di errore di predizione → minimizzazione dell'errore tramite aggiornamento del modello predittivo (logica *bayesiana*). Secondo questo approccio, l'equilibrio dinamico tra predizioni e input sensoriali permette al cervello di interagire in modo ottimale con l'ambiente, rendendo la percezione un processo profondamente dipendente dall'anticipazione e dalla correzione degli errori. Rispetto alle teorie classiche della percezione che abbiamo visto alla base delle altre teorie, la teoria del *predictive coding* offre diversi vantaggi per lo studio di fenomeni complessi come l'UVE:

- a) in primo luogo, l'abbandono dell'idea di una sequenza lineare e modulare nella percezione, a favore di un approccio distribuito e dinamico, permette di concepire la percezione come un processo integrato in cui vari livelli della gerarchia neurale operano simultaneamente per minimizzare l'errore di predizione. Il distacco dalle logiche modulariste, quindi, permette di inquadrare il fenomeno dell'UVE in un'ottica più distribuita, che coinvolge l'intera struttura neurocognitiva dell'individuo e non solamente alcuni processi che possono convergere o confliggere tra di loro.
- b) in secondo luogo, uno dei punti centrali del *predictive coding* è che il cervello non si limita a ricevere passivamente l'input sensoriale, ma formula attivamente predizioni su ciò che si aspetta di percepire. Ciò permette di riconfigurare la comprensione dell'UVE includendo altri fattori oltre all'apparenza del robot, come le sue modalità di movimento o il modo in cui si allinea con le aspettative che abbiamo nei suoi confronti: secondo gli studi a supporto della VEH che adottano la struttura interpretativa del *predictive coding*, quando interagiamo con un robot che ha caratteristiche fisiche molto simili a quelle umane, il cervello genera predizioni basate su esperienze precedenti con altri esseri umani. Tuttavia, se le caratteristiche del robot – come il movimento, le espressioni facciali o la prosodia – non corrispondono esattamente a quelle attese per un essere umano, si verifica un errore di predizione.



- c) Il *predictive coding* fornisce un quadro in cui l'UVE può essere interpretato come un *breakdown* percettivo rilevante e, dunque, può essere inserito in uno sviluppo processuale e dinamico. A differenza degli altri approcci, in cui l'UVE sembra emergere puntualmente date certe condizioni di partenza predefinite, in questo quadro il fenomeno si realizza all'interno di uno sviluppo temporale e spaziale della cognizione, che evolve in risposta alle discrepanze tra le predizioni e gli input sensoriali. Secondo questo approccio, l'UVE non è un evento statico o inevitabile, ma una rottura percettiva che emerge gradualmente attraverso continui tentativi di ridurre l'errore di predizione, suggerendo anche che l'entità e la manifestazione del disagio dipendano dalla continua rinegoziazione tra le aspettative del sistema cognitivo e le informazioni sensoriali, e che tale rottura possa essere attenuata o amplificata in base alla capacità del cervello di correggere le proprie predizioni.

Ora, a ben vedere, anche se riconfigura il modello classico della cognizione rintracciando la causa dell'UVE non al livello dell'elaborazione cognitiva ma a quello della predizione e del raffronto con i *prior*, l'approccio del *predictive coding* conserva lo stesso ordine di problematiche evidenziato per le altre teorie nella spiegazione del UVE:

- i) anche in questo caso, il problema dell'UVE è formulato a partire da un riconoscimento categoriale, in cui meccanismi percettivi di alto livello informano e influenzano quelli a livelli più bassi. Anche nel caso della VEH, infatti, il nodo della questione sembra risiedere nel modo in cui alcuni network neurali computano gli input percettivi provenienti dall'esterno, in un processo prettamente interno alla cognizione che produce una previsione riguardo a ciò che dovrebbe accadere sulla base di categorie concettuali storate nella memoria dell'individuo: l'aspetto dell'androide, così vicino a quello di un essere umano, richiama una conoscenza concettuale circa ciò che l'entità dovrebbe fare per rispettare le condizioni poste dalla sua apparenza, ma ciò si sgretola alla luce delle modalità meccaniche di impiego del movimento. Ciò che accade è che alla registrazione dell'input percettivo corrisponde un output neurocognitivo relativo alla predizione dell'errore, che viene disatteso al confronto con il comportamento realizzato. Anche per la VEH, quindi, la percezione e la cognizione sono processi temporalmente distinti, in cui la prima serve a mettere in moto la seconda anche se in un processo dinamico di aggiustamenti tra rappresentazioni e processi neurocognitivi.
- ii) ciò implica che anche la VEH, pur includendo diversi fattori nell'ipotesi, come il ruolo delle conoscenze precedenti, delle aspettative e del comportamento del robot, parte dalla *humanlikeness* come fattore scatenante dell'UVE e come banco di prova su cui le aspettative vengono riportate. Per la VEH, infatti, l'aspetto del robot costituisce il terreno su cui tutto il

resto viene confrontato e il disallineamento delle dimensioni cognitive, percettive e predittive è ciò che causa l'UVE data la rilevazione di un *mismatch* tra il grado di *humanlikeness* del robot e il dispiegarsi del suo comportamento.

La ragione di questa convergenza di problemi è facilmente rintracciabile nel fatto che i modelli del *predictive coding*, pur differendo dai precedenti e capovolgendo il tipo di spiegazione cognitiva, rimangono in ogni caso all'interno dello stesso quadro funzionalista, anche se di tipo connessionista (Kirchhoff & Kiverstein 2019; Kirchhoff & Robertson 2018; Gallagher 2017; Hutto e Myin 2017). Oltre a ciò, abbiamo già evidenziato come essa sia insufficiente a esaurire la complessità estetico-percettiva (§5.2.1.2) che genera la sensazione dell'*uncanny*, che può invece trovare la sua soluzione all'interno di un approccio teorico radicalmente differente da quello funzionalista come quello che abbiamo finora messo al centro della nostra trattazione: l'enattivismo. Si tratterà quindi di mostrare come le riflessioni sviluppate dagli approcci del *predictive coding*, una volta sottoposte a critica, possano condurci a un nuovo modello di spiegazione dell'UVE, inquadrandole in chiave sensorimotoria ed enattiva. Inizieremo a mostrare, allora, come una riformulazione enattivista sia utile a partire da uno sguardo alternativo a un concetto al centro di ognuna delle teorie e degli studi che abbiamo visto riguardare l'UVE: il concetto di *humanlikeness*.

#### 4.4.6.1 Note critiche al concetto di *human-likeness*

Nel capitolo III (§3.4.1.1), abbiamo parlato di *humanlikeness* all'interno del discorso sull'antropomorfismo. In quella sede, abbiamo analizzato come alcuni esperimenti (Phillips et al. 2018; Lupkowski et al. 2017; Ruijten et al. 2019; von Zitzewitz et al. 2013; cfr. Mahdi et al. 2022), che avevano l'intento di misurare quantitativamente la somiglianza all'umano da parte del robot, si costruissero a partire da una concezione internalista della mente e della cognizione sociale, e abbiamo ricondotto le difficoltà di stabilire parametri oggettivi di quantificazione della somiglianza all'umano a una visione specifica sulla percezione. In particolare, ciò che sottolineavamo era la fiducia, condivisa da tutti gli studi citati, di riuscire a scoprire la definizione definitiva di *humanlikeness* attraverso l'indagine delle sue caratteristiche essenziali: come sottolineato da Phillips et al. (2018: 106), «a precise metric is needed to compare different robots on a common scale and allow researchers to replicate findings with robots of equivalent human-likeness». All'identificazione di questo presupposto abbiamo fatto seguire una riflessione sulle fondamenta teoriche chiamate a giustificare la validità: i ricercatori indagano la *humanlikeness* assumendo che ne esista un concetto omogeneo, astratto, stabile e universalmente accessibile, caratterizzato da attributi semantici che permettono di individuare la presenza (o l'assenza) della qualità in oggetto in ogni artefatto o entità

incontrata. Dunque, l'ipotesi alla base degli studi consisteva nel credere che, raccogliendo stimoli percettivi simili e richiamando significati astratti associati al concetto di *humanlikeness*, come accaduto in Phillips et al. (2018), gli individui sarebbero stati in grado di discernere tra diverse entità e costruire una gerarchia di somiglianza all'umano che validasse il concetto che gli stessi ricercatori ponevano a monte dell'indagine: una visione completamente ideale del concetto, che si costituisce tramite la selezione di essenze e tratti semantici che vengono postulati essere anch'essi condivisi universalmente. In altre parole, il compito che i ricercatori chiedevano ai soggetti degli esperimenti era quello di riportare l'occorrenza a un tipo<sup>181</sup>, ovvero alla categoria di riferimento, tramite una legge che viene ipotizzata essere alla base del processo percettivo stesso. In effetti, una concezione simile non può che essere il risultato di un approccio alla percezione del tipo assunto nella maggior parte degli studi sull'UVE: nella CUH, ad esempio, la percezione funzionerebbe tramite regole di riconoscimento e categorizzazione, secondo il meccanismo modulare che abbiamo già approfondito. Se prendiamo in considerazione gli assunti epistemologici del *predictive coding*, invece, il discorso è leggermente diverso: la percezione non ha lo scopo primario di categorizzare e riconoscere lo stimolo, ma quello di anticiparlo e rendere meno probabile possibile l'insorgenza di un errore di predizione circa gli sviluppi cognitivi successivi alla presentazione dello stimolo. Tuttavia, anche qui il processo parte dalla registrazione di una caratteristica *humanlikeness* del robot e il processo di predizione si sviluppa a partire dalla rappresentazione concettuale del robot come un'entità dai tratti estetici quanto meno riconducibili alla categoria dell'umano. Di conseguenza, anche la VEH basata sul *predictive coding* parte dall'assunto che l'attività che mette in moto il processo predittivo sia, in qualche modo, il riconoscimento dell'entità come umana, fattore che richiama nella rete computazionale dei *prior* da cui verranno generate le aspettative e a cui verranno ricondotti gli esiti del processo di minimizzazione.

Nel commento proposto come interpretazione alternativa a queste teorie nel caso della *humanlikeness* abbiamo chiamato in causa una serie di teorie sviluppate da una branca della semantica cognitiva (Rosch 1975; Rosch & Lloyd 1978; Violi 1997; Lakoff 1987) che si opponevano proprio a alla visione secondo cui i concetti vengono definiti a partire dalla loro scomposizione essenziale, e proponevano un approccio dinamico del significato che si basava sull'esperienza individuale e sul contesto in cui l'interpretazione del soggetto avviene. Nel caso della critica alla nozione di *humanlikeness* posta alla base degli studi sull'antropomorfismo, questo argomento – insieme ad altre ragioni – ci era parso capace di aprire la strada a una riflessione sul ruolo del corpo e dell'esperienza incarnata nella realizzazione del fenomeno, che abbiamo realizzato in seguito (§3.4.2.1). Ora

---

<sup>181</sup> Questo processo è uno dei temi più affrontati all'interno di una semiotica della percezione, il cui studio è stato inaugurato da Umberto Eco (1997) e sviluppato sotto diversi aspetti da una visione pragmatista della cognizione e del significato (Violi 1997; 2012; Paolucci 2010).

possiamo approfondire ulteriormente cosa significhi proporre la tesi secondo cui il significato sia da considerare diversamente da quanto fatto nelle teorie funzionaliste e modulariste della cognizione, ponendo invece l'accento sulla dimensione pragmatica dei concetti astratti e vedendo come questa concezioni richiami una teoria della cognizione che pone alla base, ancora una volta, il ruolo del corpo, dell'interazione con l'ambiente e della storia dell'agente interpretante, che può risultare utile anche nell'indagine dell'UVE. Per farlo, vogliamo partire da un altro studio, che può illuminare un aspetto molto spesso lasciato in disparte negli studi sulla *humanlikeness*: la difformità di cosa somiglia all'umano.

In questo studio, Lupkowski e colleghi (2017) hanno progettato un esperimento per verificare la quanto efficacemente gli individui possano giudicare i gradi di somiglianza umana in avatar generati al computer. In particolare, i ricercatori si sono domandati se i partecipanti all'esperimento avrebbero ordinato le rappresentazioni allo stesso modo all'interno di una categoria specifica tra robot, androidi, zombie, avatar e umani (task 1). Inoltre, ai partecipanti sono state mostrate tre rappresentazioni per ciascuna categoria (fig. 5.2) e è stato richiesto loro di valutare la *humanlikeness* di ogni rappresentazione e di scegliere il personaggio più rappresentativo per ogni categoria (task 2).



Fig. 5.5 (Lupkowski et al. 2017: 127)

I risultati, però, hanno completamente sconfessato le ipotesi dei ricercatori, dal momento che né nel primo, né nel secondo studio si sono registrati risultati omogenei. Infatti, nel primo *task* i partecipanti non hanno collocato gli oggetti su una scala condivisa di crescente somiglianza umana, come invece i ricercatori avevano inizialmente ipotizzato: ciascun partecipante ha elaborato una diversa scala soggettiva basata sulla propria nozione personale di *humanlikeness*. Analogamente, nel

secondo studio non è emerso alcun consenso circa il personaggio più rappresentativo per ogni categoria, rivelando un approccio individuale e idiosincratico alla percezione della *humanlikeness*. Questi risultati hanno portato i ricercatori a concludere che tali indagini debbano essere parte integrante dello studio dell'UVE, poiché non sembra possibile stabilire una scala oggettiva basata su parametri standardizzati attraverso cui collocare l'aspetto di un robot sulla scala della somiglianza all'umano. La valutazione della *humanlikeness*, quindi, dovrebbe essere adattata ad ogni singolo studio.

Da questo studio, ma anche da quello condotto da Cheetham e colleghi (2011) che abbiamo analizzato in precedenza e da altri incontrati nel corso della trattazione (Phillips et al. 2018; Bartneck et al. 2009; Ellis et al. 2005; Hinds et al. 2004; Kiesler & Goetz 2002; Kim et al. 2014; von der Pütten & Kramer 2012), possiamo ricavare una serie di considerazioni relative alla difficoltà di stabilire una dimensione oggettiva e universale per la *humanlikeness*, dato che i tentativi che abbiamo incontrato fino a questo momento sono tutti incappati nel bisogno di definire stabilmente cosa la somiglianza all'umano sia e quali siano le caratteristiche che la rendono riconoscibile. Questa operazione è resa necessaria dall'impostazione funzionalista e cognitivista alla base delle teorie impiegate nella spiegazione dell'UVE, che dà però per scontato che la caratteristica della somiglianza all'umano possa essere assunta a priori dell'indagine o, in alternativa, indagata con mezzi appartenenti al dominio del logico, dell'astratto e del ragionamento compositivo. Questi studi testimoniano che riportare le occorrenze alle categorie universali risulta un compito molto meno semplice e universale di quello che si credeva e che, probabilmente, non è questo il meccanismo che contraddistingue il processo percettivo nella cognizione umana. In particolare, lo studio di Lupkowski e colleghi illumina un aspetto che fino ad ora era rimasto sottotraccia: ogni individuo ha la propria idea, esperienza, concetto e percezione della *humanlikeness*, e questa deve essere valutata individualmente per stabilire una correlazione tra questa variabile e la valenza emotiva suscitata, come nel caso dell'UVE.

È interessante notare, tuttavia, che proprio questa individualità del concetto, questa proprietà esclusivamente personale, costituisce ciò che abbiamo in comune e che ci permette di riconoscere qualcosa, a prescindere dal livello di somiglianza presentato, come simile a qualcos'altro. Infatti, l'incapacità di rintracciare una definizione universale per il concetto di *humanlikeness* non implica che sia il caos, l'opinione personale o la pura intuizione a governare in ogni aspetto del processo di riconoscimento del robot come somigliante all'umano. Sebbene ogni individuo possa avere una concezione unica di ciò che implica la *humanlikeness*, o di come dovrebbe apparire un'entità simile a un essere umano, è probabile che vi sia un accordo sul fatto che la semplice sedia su cui sono seduta ora abbia un livello di somiglianza umana molto basso, se non addirittura inesistente, mentre i robot raffigurati nelle immagini precedenti presentano un grado considerevole di somiglianza all'umano.

La situazione diventa più complessa quando si deve scegliere tra diversi volti robotici, e la difficoltà aumenta con il numero di robot. Ciò avviene perché il significato viene attribuito attraverso stabilizzazioni pragmatiche e regolarità esperienziali, piuttosto che tramite categorie concettuali che fissano regole definitive e logicamente descrivibili.

La categoria concettuale su cui le teorie della cognizione funzionalista costruiscono il proprio approccio alla percezione e alla *humanlikeness*, allora, diventa qualcosa che può essere individuato solamente a posteriori, come regolarità realizzata all'interno dell'insieme presentato (sia al livello dell'esperienza dell'individuo, sia nel dominio locale delle rappresentazioni presentate), e che non può fungere da elemento guida dell'indagine sulla determinazione del suo significato. In questo modo, diventa ancora più chiaro il motivo per cui i partecipanti all'esperimento di Lupkowski et al. (2017) hanno stabilito ognuno una propria classifica di robot somiglianti all'umano e perché abbiano espresso scelte differenti nell'individuazione della rappresentazione prototipica; allo stesso tempo, si chiarisce anche perché, nell'esperimento di Cheetham e colleghi (2011), lo sforzo cognitivo non debba necessariamente corrispondere al confine tra categorie, dal momento che i confini di ognuna per ciascuno sono mobili e permeabili, infuse dall'esperienza che ognuno ha con quel determinato concetto di cui si richiama l'utilità. Se ci aspettiamo che la cognizione sia confinata all'interno dell'individuo e che sia identificabile esclusivamente nella sua forma logica, ci troveremo a dover giustificare perché al confine tra le categorie concettuali non emerge alcun UVE e saremo sorpresi quando lo sforzo cognitivo maggiore viene riscontrato nei pressi della categoria dell'umano, e non coincidente con il (presunto) confine di categoria. Gli esseri umani possiedono forme diverse e alternative di comprensione, che vanno oltre la logica, la semantica compositiva e le regole di associazione; esiste una dimensione che trascende logica e semantica, legata agli aspetti estetici, incarnati e pragmatici della cognizione: vale a dire, l'esperienza.

Questa dimensione non si fonda su occorrenze categoriali, né sul riconoscimento tramite regole, ma su logiche incarnate che permeano le interazioni che un agente istituisce con il proprio ambiente, con la rete delle possibili azioni che può intraprendere in quel preciso *micromondo* (Varela 1992b). Questa è la dimensione da indagare se vogliamo capire come sia possibile che la somiglianza del robot all'umano sia un aspetto fondamentale della HRI, senza ricorrere a spiegazioni razionalistiche o modulariste che comportano i problemi che abbiamo riscontrato, ma che si concentra sulle modalità di aggiustamento individuali dell'agente con il proprio ambiente e con le entità che lo popolano, sviluppandosi nel processo continuo di conciliazione tra quello che potrebbe essere e quello che effettivamente è.

Possiamo provare, quindi, a seguire i tratti di una teoria incarnata della cognizione e delle proposte in merito alla coordinazione di aspettative future e presupposti abituali incarnati considerando la

proposta enattivista. L'enattivismo, come teoria che mette al centro il ruolo del corpo, dell'azione e dell'esperienza incarnata, ha già proposto alcune soluzioni alternative sia all'approccio del *predictive coding*, che abbiamo visto essere un percorso promettente ma aprire lo stesso ordine di problemi implicato anche nelle altre spiegazioni, sia alla necessità di comprendere l'emersione di significati percettivi in termini esperienziali ed ecologici. Questo perché, per l'enattivismo, percepire non è apporre etichette sui percetti, ma un'azione guidata dalla percezione costitutiva delle capacità cognitive dell'individuo.

#### 4.4.7 Percezione è azione, azione è cognizione: significati incarnati e pratiche percettive

Nel corso della nostra trattazione, ci siamo più volte trovati a ribadire le tesi a fondamento del paradigma enattivista per evidenziare la possibilità di intraprendere un'altra via nel tentativo di dare conto di fenomeni centrali sia nello studio della cognizione, sia della HRI. Nel capitolo §2, abbiamo visto come una nuova generazione di robot basata su alcuni dei presupposti enattivisti abbia rivoluzionato il campo della robotica aprendo a possibilità fino a quel momento precluse; nel capitolo §3, una teoria enattivista della cognizione sociale ci ha permesso di rileggere alcuni dei fenomeni centrali per la HRI, come l'antropomorfismo e l'*embodiment*, attraverso lo sguardo dell'osservatore e di allargare la comprensione del fenomeno non solo a meccanismi cognitivi interni ma all'interazione situata degli agenti coinvolti; nel capitolo §4, abbiamo adottato un approccio enattivista per chiarire come sia possibile rileggere il ruolo dell'intenzionalità nella HRI da una prospettiva incarnata e legata alla capacità del robot di realizzare forme sincretiche di comportamento. Anche per comprendere come sia possibile che le spiegazioni all'UVE abbiano, fino a questo momento, fornito analisi solamente parziali del fenomeno, ritornare sui principi cardine della teoria enattivista della mente può riuscire a gettare una nuova luce sui fenomeni che stiamo analizzando e sciogliere il nodo che ha finora tenuto insieme i presupposti cognitivisti e la caratterizzazione dell'UVE. Nel tentativo di spiegare l'UVE, infatti, tutte le analisi si sono limitate a segmentare e semplificare il fenomeno, interpretandolo come un errore di categorizzazione percettiva o un deficit di elaborazione cognitiva. In questo quadro frammentato e controverso, il paradigma enattivista ci permette di cogliere l'interdipendenza tra percezione, azione e cognizione, mettendo in luce come tali facoltà emergano non come eventi separati, ma come manifestazioni di un'unica attività di *sense-making* incarnato dell'agente (Varela et al. 1991; Di Paolo et al. 2010; Thompson 2007; Gallagher 2017, 2021). Attraverso questo sguardo, mostreremo come sia possibile accogliere, se rilette, tutte le spiegazioni fin qui annoverate sotto un approccio che privilegia il significato pragmatico, l'azione incarnata, la percezione ecologica e la cognizione enattiva.

Seguendo l'approccio enattivista, allora, ripensare al processo attraverso cui si arriva a riconoscere qualcosa come tale significa istituire un legame tra le capacità cognitive dell'agente, date dall'accoppiamento dinamico dell'organismo con l'ambiente, e i contesti di realizzazione dell'azione. Questo assunto proviene direttamente dalle prime formulazioni dell'approccio enattivista, che si costituiscono proprio a partire da una critica serrata agli approcci alla percezione del tipo funzionalista che abbiamo analizzato nei paragrafi precedenti: «cognition [is] not a recovery or projection but embodied action» (Varela et al. 1991: 172). Nel definire il concetto di *azione incarnata*, Varela, Thompson e Rosch spostano l'attenzione dai meccanismi di raccolta e processazione delle informazioni nel livello computazionale al modo in cui l'organismo realizza le sue stesse possibilità di intervento in un circolo percettivo-agentivo che costituisce – e non segue o precede – la capacità cognitiva dell'agente (cfr. anche Gangopadhyay & Kiverstein 2009: 64): la cognizione è radicata nel corpo perché profondamente dipendente dal tipo di esperienza che emerge dalla conformazione dell'*embodiment* dell'organismo, ovvero dalle capacità sensorimotorie che il corpo consente e che contribuisce a strutturare dato l'accoppiamento strutturale con l'ambiente, il quale si realizza in *affordance* biologici, sociali e culturali; la cognizione è azione proprio perché i processi percettivi e agentivi, la percezione e l'azione, sono inseparabili e costituiscono la capacità cognitiva dell'agente dall'interno, in un processo dinamico di mutua evoluzione delle capacità che si influenzano vicendevolmente.

Il nodo tra percezione, azione e cognizione trova la sua più completa realizzazione proprio nella nozione di *enaction*, ovvero il fatto che «(1) perception consists in perceptually guided action and (2) cognitive structures emerge from the recurrent sensorimotor patterns that enable action to be perceptually guided» (Varela et al. 1991: 173). Nella formulazione enattivista, percezione, cognizione e azione non vengono più interpretati come meccanismi autonomi del sistema cognitivo, da analizzare a considerare separatamente uno dell'altro, ma come gli elementi discretizzati di un ciclo di produzione di senso che permettono una presa scientifica per analizzare il funzionamento della mente umana. La percezione si configura come un processo attivo e agentivo di esplorazione del mondo, un insieme di abilità che coinvolgono l'intero organismo, condizionate dal modo in cui l'agente vive il suo ambiente e che si manifestano in realizzazioni pratiche di azioni sensorimotorie. Uno dei risultati emergenti dal loop percettivo-agentivo realizzato dall'agente è la sua capacità cognitiva, che scaturisce dalle possibilità di azione e, al contempo, le permette.

In sostanza, percepire qualcosa significa enattivare significati pragmatici, ovvero legati alle possibilità di azione che l'organismo individua nell'ambiente. All'interno di questa rivoluzione dello sguardo alle dinamiche cognitive, anche il processo di categorizzazione e riconoscimento è visto in maniera molto diversa: la categorizzazione diventa un'attività di *meaning making* (Di Paolo et al.



2010; 2018), di costruzione del significato che consta di enattivazioni individuali e convocazioni della collettività, in cui, nella moltitudine di significati enattivati dall'apertura del loop agentivo-percettivo all'oggetto, si attenziona una porzione riconoscibile, i cui legami con l'esperienza e *l'embodiment* dell'agente stabiliscono il rispetto sotto cui le possibilità di azione possono venire realizzate:

One of the most fundamental cognitive activities that all organisms perform is categorization. By this means the uniqueness of each experience is transformed into the more limited set of learned, meaningful categories to which humans and other organisms respond. [...] The basic level of categorization, thus, appears to be the point at which cognition and environment become simultaneously enacted. The object appears to the perceiver as affording certain kinds of interactions, and the perceiver uses the objects with his body and mind in the afforded manner. Form and function, normally investigated as opposed properties, are aspects of the same process, and organisms are highly sensitive to their coordination. And the activities performed by the perceiver/actor with basic-level objects are part of the cultural, consensually validated forms of life of the community in which the human and the object are situated – they are basic-level activities. (Varela et al. 1991: 177).

Se applichiamo questa visione al processo di riconoscimento del robot antropomorfo, individuare una somiglianza tra le sue sembianze e l'essere umano non vuol dire primariamente riportare il percepito al concetto, ma individuare il *micromondo* enattivato dalle *affordance* presentate tramite la componente mimetica del robot. Come abbiamo visto in §5.1.1, le modalità attraverso cui Varela identifica il micromondo, all'interno del quale l'agente intreccia legami interazionali e attentivi con alcune delle sue parti, si configurano nella rilevazione – nella *detection* (cfr. 4.3) – di caratteristiche che emergono come salienti nell'ecologia distribuita della continuità del *sense-making*. I fattori che avevamo rintracciato come identificativi del micromondo per come descritto da Varela (1992b) consistevano i) nella loro *ricorsività* di emersione, legata alle occorrenze di realizzazione di una situazione interazionale nella storia di enattivazioni esperienziali dell'agente e ii) nel condizionamento operato dalle dinamiche sensorimotorie realizzate dall'agente, le quali contribuiscono sia a costruire il micromondo come termine *a quo* insito nell'*embodiment*, sia a realizzarlo come risultato *ad quem* della cognizione. Nel circolo dinamico di percezione e azione, l'emersione del micromondo che il robot contribuisce a enattivare consiste nelle possibilità delle azioni che possiamo svolgere con lui: la sua somiglianza ad altri umani che abbiamo incontrato nel corso della nostra storia interazionale si inserisce all'interno della costruzione continua di significato, presentandosi come un elemento di attenzione e acquistando senso all'interno del contesto in cui l'occasione si presenta. In altre parole, insieme agli occhi del robot, cogliamo la possibilità di essere oggetto del suo sguardo; insieme alle sue mani, anticipiamo la possibilità di stringerle; insieme alla sua espressione facciale, sentiamo la possibilità di comprenderne l'emozione.

In questo senso, la *humanlikeness* è, sì, una qualità espressa dalle sembianze del robot, ma è rilevante solamente se inserita all'interno del processo di continui aggiustamenti materiali e

intersoggettivi che ci troviamo costantemente ad allacciare e sciogliere nel passaggio tra micromondi o all'interno dello svolgimento dello stesso. È interessante notare che recentemente, a partire da questo sguardo sulla percezione come strettamente legata alle possibilità di azione e intimamente correlata alle modalità cognitive dell'individuo, sono state evidenziate nuove possibilità di definizione del processo percettivo, non più legate solamente all'individuo in rapporto alle capacità interazionali mostrate con il proprio ambiente, ma connotate da una intrinseca socialità: quel legame tra persone che contraddistingue l'intera vita dell'essere umano e che condiziona anche le modalità attraverso cui l'individuo percepisce il proprio ambiente.

In particolare, in un bellissimo articolo, Arango (2019) mette in luce, attraverso il concetto di *vedere-come* elaborato da Wittgenstein, che percepire è un'attività che viene sempre realizzata in un contesto sociale, concretizzata in uno sviluppo pragmatico e incarnata negli abiti sedimentati della nostra esperienza:

[...] perception is always perceiving-aspects. This insight yields a substantive set of claims: that in perception we always perceive things aspectually, that sensory input alone is insufficient to account for this aspectual nature, that this aspectual nature is oftentimes experientially hidden, and that this aspect-grasping has an element of “familiarity” or “recognition” that is not a thinking or a judging, but rather a certain bodily engagement. Seeing is always seeing-as. This point is supplemented by a second one about the relationship between aspects and doings. (Arango 2019: 39)

L'autore procede a sviluppare un concetto interessante di percezione legata all'azione, che prende il nome di *perceptual practice*: «the enactment of culturally structured, normatively rich techniques of interaction with multi- and intermodal perceptible material. Perception is woven in a type of social practices [...]» (Ibid.). Queste pratiche percettive mostrano una serie di caratteristiche che possono risultare utili alla nostra indagine sulle modalità di riconfigurazione del concetto di comprensione nell'attività dell'UVE, soprattutto in relazione ad uno sguardo sulla *humanlikeness*.

Innanzitutto, le pratiche percettive sono caratterizzate dalla *contestualità*, dal momento che per l'essere umano la determinazione dei significati percettivi, cioè di quello che emerge come l'oggetto della percezione, è rintracciabile solamente nel novero delle pratiche concrete che si realizzano nel caso di specifiche circostanze sociali, le quali includono necessariamente anche l'ambiente fisico e instradano le modalità di percepirlo; sono *performative*, perché si costituiscono di azioni, reazioni e disposizioni correlate alle percezioni; sono *espressive*, perché attraverso di loro si esprime la forma di vita di una comunità nella loro interazione con il mondo guidata dalle sensazioni e attività corporee; sono *normative* nel loro essere valutabili attraverso criteri di intelligibilità e concretezza pragmatiche – ovvero nella loro efficacia in relazione al contesto e nel loro significato condiviso e situato –, criteri radicati nei processi attraverso cui i membri di una comunità acquisiscono modi di interagire con il proprio ambiente e con gli agenti che lo popolano (Arango 2019: 42).

Il legame di percezione e azioni veicolata dal concetto di pratiche percettive, allora, può aiutarci a comprendere come anche la *humanlikeness*, intesa come qualità attribuita all'aspetto del robot, sia il frutto di un insieme di contingenze sensorimotorie, sociali, culturali e pragmatiche che si realizzano nell'esperienza e nel comportamento dell'individuo:

- L'aspetto contestuale ci consente di vedere come la somiglianza all'umano esibita dal robot non sia una qualità intrinseca del suo aspetto, ma possa variare in accordo a fattori di tipo sociale, culturale e ambientale in cui l'incontro si verifica. Anche nel caso di esperimenti con un basso, se non nullo, grado di contestualizzazione del robot, come quelli svolti tramite l'utilizzo delle *morphed image* o delle immagini virtuali, la percezione della *humanlikeness* del robot può variare in base a fattori circostanziali della presentazione: le disposizioni affettive e interazionali momentanee dell'individuo, l'inclinazione caratteriale, l'educazione, l'esperienza con altri artefatti dello stesso tipo, i mezzi disponibili all'individuo per inserire quell'incontro all'interno del quadro più ampio della sua esperienza di vita, ecc. Il contesto, in questo senso, si manifesta su più livelli e, dal momento che la facoltà cognitiva dell'individuo viene esercitata sempre in una circostanza realizzata e mai ideale, il tentativo di rendere asettico il *setting* sperimentale non presuppone l'annullamento della contestualizzazione.
- La performatività della pratica percettiva mostra che la percezione della *humanlikeness* è un processo attivo e incarnato, che si sviluppa a partire dai significati enattivati in relazione alla percezione del robot come agente. Non si tratta semplicemente di riconoscere delle caratteristiche visive, che il robot mostrerebbe impresse sul suo volto, ma di un coinvolgimento attivo in cui l'agente umano partecipa a una serie di azioni e reazioni che strutturano la sua esperienza percettiva, innescate in un panorama di significati. In altre parole, la *humanlikeness* non è una caratteristica fissa, statica, riscontrabile oggettivamente in ogni sua manifestazione e che può essere ascritta ad ogni robot in un continuum di artefatti somiglianti all'umano, ma un fenomeno performativo che si costruisce nel corso della percezione e che si dispiega nelle modalità di interazione, anche immaginata: quanto più il robot riesce a rispondere in modo appropriato e coerente ai significati pragmatici e sociali enattivati dell'agente umano, tanto più sarà percepito come *humanlike*, generando reazioni antropomorfizzanti (Abubashait & Wiese 2017; Kiesler et al. 2008; Mishra et al. 2022).
- L'espressività si manifesta nelle diverse qualità che la *humanlikeness* assume all'interno di contesti culturali specifici, poiché le pratiche percettive che regolano il riconoscimento di qualcosa come simile all'umano sono profondamente influenzate dalle forme di vita collettive e condivise (Papadopoulos & Koulouglioti 2018; Roselli et al. 2022; Castelo & Sarvary 2022). Questo aspetto rinforza l'idea che la percezione della *humanlikeness* non sia un'attività

universalmente realizzata allo stesso modo da tutti gli individui umani, ma un'azione che viene modellata dalle norme e dai valori che ogni cultura attribuisce al concetto di "umano". Questo significa che i criteri con cui si stabilisce se un robot è *humanlike* possono variare sensibilmente: in alcune culture, un robot potrebbe essere considerato somigliante all'umano se evoca un certo tipo di interazione rituale o simbolica, anche se il suo aspetto fisico non è strettamente antropomorfo (Robertson 2018; Bartneck et al. 2009); in altre, invece, l'enfasi sulla somiglianza fisica e comportamentale si rivela come una caratteristica fondamentale per riconoscere la *humanlikeness* (Nomura et al. 2006; MacDorman 2006), come accade nella nostra impostazione culturale occidentale, ma può anch'essa essere soggetta a variazioni (Green et al. 2008). Pertanto, la qualità della *humanlikeness* non è solo il prodotto di una valutazione individuale, ma si sviluppa all'interno di pratiche espressive collettive che riflettono i valori culturali e sociali legati all'umano e all'umanoide. Questo aspetto della percezione è supportato anche da una serie di studi (Chua et al. 2005; Masuda & Nisbett 2001; Winerman 2006) che rintracciano differenze sulle modalità attraverso cui si realizza la percezione in individui provenienti dagli USA e altri dall'Asia orientale. Ad esempio, Chua e colleghi (2005) hanno misurato il movimento oculare in individui statunitensi e cinesi mentre osservavano una scena in cui erano presenti alcuni oggetti che si muovevano su uno sfondo, osservando che «Americans fixated more on focal objects than did the Chinese, and the Americans tended to look at the focal object more quickly. In addition, Chinese made more saccades to the background than did Americans» (2005: 12629). Allo stesso modo, è stato riscontrato che, alla presentazione della raffigurazione di un acquario contenente diversi pesci e alcuni oggetti che popolavano lo sfondo, nel ricordare il contenuto della rappresentazione gli individui americani si sono concentrati principalmente sui pesci, mentre individui provenienti dal Giappone tendevano a descrivere la scena visiva nel suo complesso, riportando anche molti dettagli riguardo alla composizione dello sfondo (Masuda & Nisbett 2001; Winerman 2006). Come sottolineato da Arango (2019: 35):

What these scientific studies show is that basic aspects of visual experience can be culturally constituted and cannot hence be assumed to be universal [...]. They also show that the content of perception is not simply a mirroring of external states of affairs but a take on them. [...]. The specificity of the experience cannot be pinned down then to the way the world is, since the way the world is gives rise to different experiences, none of which is incorrect. It has got to reside, then, at the level of aboutness – that which the experience is about – which is not only a matter of worldly reference.

- La normatività della pratica percettiva legata alla *humanlikeness* è realizzata nelle modalità di emergenza e identificazione del fenomeno. In questo senso, la *humanlikeness* è soggetta a criteri di intelligibilità sociale e pragmatica che possono variare a seconda del contesto, della

disposizione affettiva ed esperienziale del soggetto e alle modalità in cui si realizza l'azione stessa. I criteri normativi della pratica percettiva si esprimono, allora, sotto forma di *aspettative* riguardo l'interazione tra umano e robot, e stabiliscono le disposizioni attraverso cui il comportamento e l'aspetto del robot vengono percepiti e interpretati. In altre parole, l'assetto concreto della pratica percettiva realizzata di costruisce a partire dalle norme condivise e sociali, e viene modellato dalle possibilità pragmatiche di intervento e coordinazione motoria.

Riassumendo, la *humanlikeness* è una qualità che emerge dall'esperienza dell'individuo, che include quella della socialità e dell'intersoggettività, e consente il riconoscimento di oggetti, volti, persone, concetti, emozioni e sensazioni come simili a quelle umane; a sua volta, essa viene plasmata dal modo in cui l'esperienza si manifesta, non soltanto a livello concettuale, ma primariamente a livello corporeo.

È fondamentale osservare come le conseguenze di questa rilettura del concetto di somiglianza all'umano impattino profondamente sulle possibilità di comprensione del fenomeno dell'UVE. Da una parte, infatti, sottolineando la componente agentiva e sensorimotoria del processo di comprensione del significato percettivo, l'UVE passa dall'essere un fenomeno descrivibile alternativamente nel dominio del percettivo o del cognitivo, a chiamare in causa tutto l'apparato cognitivo-percettivo, sensorimotorio ed esperienziale dell'individuo: il fenomeno non è più descrivibile facendo appello esclusivamente a meccanismi funzionali interni alla processazione dello stimolo, ma deve comprendere la pluralità e la multidimensionalità dei fattori che abbiamo visto costituire le modalità di costruzione di senso dell'individuo. Esperienza, corporeità, agentività, storia e disposizioni degli agenti sono gli aspetti che costruiscono il micromondo in cui l'UVE si manifesta. Dall'altra parte, l'accento sulla costitutività della socialità e della intersoggettività della pratica percettiva incarnata ci presenta la necessità di inquadrare l'UVE sotto l'aspetto dell'interazione sociale: in questo senso, l'UVE non rappresenta solo una risposta individuale a uno stimolo ambiguo, difficilmente classificabile o incoerente, ma si configura come un fenomeno emergente nelle dinamiche sociali in cui il robot viene inserito, richiamando aspettative condivise all'interno dell'allestimento di possibilità di azione e interazione incarnate nelle possibilità agentive dell'individuo. La percezione del robot come quasi umano, ma non completamente umano, si scontra non solo con un'esperienza individuale, ma anche con i significati socialmente negoziati e culturalmente specifici che modellano le aspettative sull'interazione sociale con agenti umanoidi.

Infine, sia l'aspetto normativo, sia la nozione di percezione come *perceptually guided action* ci spingono a includere un fattore che fino ad ora abbiamo considerato solo parzialmente, ma che risulterà fondamentale per riuscire a spiegare quale sia la struttura dell'UVE nel quadro di una teoria

enattivista: le aspettative. Nelle spiegazioni fornite in §5.2.1.1-5.2.1.2, le aspettative sono state considerate, in qualche modo, inserite nel novero dei fattori che sembrano causare l'UVE: nella MPH, ad esempio, l'innescò del modulo della ToM da parte della *humanlikeness* del robot è considerato richiamare una serie di aspettative in termini di cognizione sociale; lo stesso accade nella DH, che allo stesso tempo però pone l'accento sulle aspettative comportamentali del robot. Cionondimeno, per queste due ipotesi esplicative non è il raffronto con le aspettative che causa l'UVE, ma fenomeni che dipendono dalla disattesa di queste. È nella VEH, d'altra parte, che le aspettative assumono il ruolo di fattore causale preminente dell'UVE: le aspettative create dai *prior* percettivi si trovano in conflitto con lo stimolo in entrata dato dal movimento meccanico del robot, e il processo di minimizzazione dell'errore registra un *mismatch* di cui si deve dare conto. È proprio in relazione a questo tipo di spiegazioni e in linea con l'adesione alle teorie del *predictive coding* che le aspettative assumono l'importanza maggiore, dal momento che stabiliscono il banco di prova su cui la minimizzazione dell'errore di predizione avviene.

Si potrebbe pensare, allora, di poter abbracciare senza grandi riserve il paradigma suggerito dalla VEH, dal momento che i) mette in luce l'importanza del movimento del robot; ii) adotta delle modalità sperimentali che vedono l'agente robotico muoversi nel suo ambiente; iii) si discosta dalle concezioni modulariste della percezione e ristruttura la spiegazione su un processo dinamico di elaborazione dello stimolo percettivo. Tuttavia, come approfondiremo meglio nel prossimo paragrafo, alcune importanti differenze tra il paradigma teorico del *predictive coding* e una visione enattivista della cognizione sono state messe in luce in diverse sedi (Gallagher & Bower 2013; Gallagher & Allen 2018; Kirchhoff & Robertson 2018; Kirchhoff (Ed.) 2018; Bizzari & Hipolito 2016; Hutto & Hipolito 2024). Possiamo suggerire preliminarmente una sostanziale differenza, che per alcuni si configura come un segno di incompatibilità<sup>182</sup> (Gallagher 2023; Di Paolo 2014; Di Paolo et al. 2021;

---

<sup>182</sup> Le maggiori discrepanze tra l'approccio neurocognitivo basato sulle capacità funzionali predittive del cervello e l'approccio enattivista riguarda principalmente il concetto di *active inference* per come formulato da Friston (2005, 2010). Recentemente, Di Paolo, Thompson e Beer (2021) hanno fornito un'estensiva disamina dei motivi per cui il principio dell'inferenza attiva alla base del *Free Energy Principle* (FEP) – di cui la teoria del *predictive coding* è una declinazione computazionalista – sia incompatibile con la teoria enattivista, anche se entrambi gli approcci fanno affidamento alla teoria dell'autopoiesi per come sviluppata da Maturana e Varela (1980): secondo Di Paolo et al. (2021), la chiave per comprendere l'incompatibilità degli approcci risiede nella differenza posta da Maturana e Varela tra *struttura* e *organizzazione* (Maturana e Varela 1980; cfr. §2.3.1.2). Allo stesso tempo, negli ultimi anni sono comparsi diversi contributi a favore di un'integrazione dell'approccio basato sul FEP con l'enattivismo, promossa allo scopo di ampliare la portata della teoria enattivista della cognizione (Kirchhoff 2015; Kirchhoff et al. 2018; Ramstead et al. 2018) o di integrare quest'ultima nel più ampio quadro della teoria fisica di Friston (Allen & Friston 2018; Bruinenberg et al. 2018; Kiverstein 2020; Gallagher & Allen 2018). Entrare nel dibattito contemporaneo sulla compatibilità o inconciliabilità di questi modelli teorici va oltre l'intento della presente trattazione. Possiamo comunque procedere con le nostre riflessioni in merito al ruolo svolto dalle anticipazioni percettive e dalla creazione di aspettative lasciando da parte il dibattito epistemologico relativo alle radici teoriche e concentrandoci sulle declinazioni che le teorie predittive hanno assunto nello specifico, come quella del *predictive coding* nella prospettiva computazionalista e quella del *predictive engagement* in ottica enattivista. Tale operazione è resa possibile dalle differenze sostanziali che i due approcci producono in termini di visione generale del processo cognitivo: come verrà approfondito a breve, l'approccio enattivista consente di considerare una pluralità di fattori che altrimenti resterebbero esclusi (cfr. §5.3.1).

Gärtner & Clowes 2017), tra l'approccio del *predictive coding* alla base della VEH e l'approccio enattivista che stiamo proponendo, semplicemente prendendo in considerazione la natura ecologica e distribuita del processo cognitivo suggerito dalla nozione di percezione come *perceptually guided action*: come abbiamo visto, nell'approccio della VEH la percezione è condizionata dal movimento dell'oggetto coinvolto, fattore considerato l'origine dell'UVE, ma solamente perché la sua rappresentazione all'interno dell'architettura funzionale confligge con le predizioni operate a livello alto dal richiamo dei *prior* innescato dalla *humanlikeness* del robot. In questo senso, come abbiamo notato, la spiegazione non si discosta di molto dalle altre che pensano all'UVE come generata da conflitti cognitivi interni al processo percettivo. Il movimento, in questi casi, è pertinente solamente in quanto rappresentato all'interno dell'operazione computazionale operata dal sistema di reti neurali che compongono la struttura cognitiva dell'individuo; l'azione realizzata della percezione non è distribuita nell'ambiente e enattivata in relazione alla contestualità e alla pragmatica dell'occorrenza pratica, ma inquadrata nel processo predittivo della minimizzazione dell'errore: un processo questo tutto interno alla mente dell'individuo, confinata e a volte sovrapponibile (Howhy 2013; Clark 2016) con la struttura neurale che la realizza, anche se condizionabile da fattori esterni provenienti dal mondo e dalle sensazioni corporee. In sostanza, se seguiamo la teoria del *predictive coding*, ciò che conta per comprendere le cause e la struttura dell'UVE è ancora una volta da rintracciare nella struttura funzionale del cervello, in un disallineamento di previsioni e realizzazioni che si riscontra al momento della processazione cognitiva dello stimolo percettivo del robot. Questo, come abbiamo visto, sembra essere uno sguardo che limita la comprensione del fenomeno, non includendo gli elementi sensorimotori, contestuali, sociali e distribuiti dell'esperienza di fronte al robot.

L'esempio che abbiamo fatto in precedenza, in cui ci troviamo di fronte a una persona che ancora non conosciamo, può aiutarci a chiarire la differenza tra i diversi approcci. Abbiamo già descritto la situazione: ci troviamo a una conferenza e qualcuno ci sta per presentare una persona. La situazione formale in cui ci troviamo a interagire comporta un'etichetta precisa, che tendenzialmente conosciamo. Nel caso in questione, quando ci presentiamo, il gesto automatico e previsto dalla circostanza e dalla presenza dell'interlocutore sconosciuto è quello di porgere la mano, affinché l'altro la stringa. Per qualche motivo, però, la persona davanti a noi non afferra la nostra. Ora, adottando l'approccio classico delle spiegazioni che abbiamo fornito fin qui, abbiamo due alternative per spiegare il fenomeno: secondo gli approcci della percezione categoriale, la situazione della mancata stretta di mano sarebbe causata dal rimando a una categoria concettuale errata. Nella interpretazione suggerita dai principi di questa ipotesi, l'errore potrebbe essere causato dalla nostra scarsa attenzione alla situazione sociale: ad esempio, potremmo trovarci di fronte a un interlocutore con le mani

impegnate, con ridotte capacità di movimento, o intento a conversare con altri colleghi. Tale spiegazione sembra più che ragionevole, ma, a ben vedere, non sembra rendere giustizia a quello che realmente accade, perché riduce l'intero evento a un problema di riconoscimento di pattern percettivi senza considerare le dinamiche agentive e incarnate dell'interazione, tralasciando quindi la complessità del contesto e dell'immediatezza dell'azione. Abbiamo compreso che un impianto teorico come quello della percezione categoriale non riesce a rendere conto dei fenomeni complessi che si sviluppano nella percezione, soprattutto se considerata all'interno di pratiche sociali e intersoggettive. D'altra parte, seguendo l'approccio del *predictive coding*, la circostanza è descrivibile in termini di *mismatch* tra le aspettative predittive e l'effettiva risposta dell'interlocutore: il nostro cervello avrebbe previsto, basandosi su modelli sociali appresi, che la mano sarebbe stata stretta. Quando questa previsione non viene confermata, si genera un errore di predizione, il quale richiede un aggiornamento del modello mentale per spiegare l'evento. Si può riconoscere che anche questo approccio soffre di una certa parzialità rispetto alla complessità delle dinamiche interazionali che abbiamo visto essere descritte da un approccio enattivista: la spiegazione resta confinata entro il dominio della cognizione individuale, centrata sull'aggiornamento di modelli interni e, sebbene includa un meccanismo di adattamento dinamico, l'approccio del *predictive coding* trascura l'importanza del coinvolgimento corporeo e intersoggettivo che caratterizza l'interazione in tempo reale. Inoltre, l'enfasi sull'errore predittivo ignora la fluidità e la creatività che caratterizzano le interazioni incarnate e contestuali, riducendo l'interazione a un problema computazionale di inferenza e aggiornamento di ipotesi. Al contrario, se adottiamo la prospettiva enattivista che vede la percezione come un'azione incarnata, il processo sarà descritto in termini ben diversi: il fallimento del gesto non è riducibile a un errore di categorizzazione o di predizione, ma è visto come una rottura del flusso di interazione incarnata tra gli agenti. L'atto di porgere la mano non è semplicemente un segnale sociale da interpretare, ma una pratica percettiva incarnata che dipende dal coinvolgimento sensorimotorio, dal contesto di realizzazione, dalle possibilità pragmatiche di intervento e concatenazione e dalle modalità abituali con cui si realizza il gesto. Quando la stretta di mano non avviene, questo non rappresenta solo un errore percettivo o cognitivo, ma un *breakdown* nell'azione reciproca, che modifica in tempo reale le possibilità di azione dell'agente e che a sua volta comporta l'enattivazione di micromondi attraverso cui cercare di compensare la rottura.

Il vantaggio dell'approccio enattivista risiede nella sua capacità di spiegare il fenomeno in modo olistico, integrando l'esperienza sensomotoria e il contesto sociale, e riconoscendo la partecipazione attiva dell'individuo nella costruzione del significato, soprattutto in relazione alle interazioni con altri individui. Le teorie del *predictive coding* alla base della VEH hanno aperto la strada ad una possibile riconfigurazione delle modalità di considerare ciò che avviene nell'UVE, introducendo fattori come



le aspettative dell'essere umano e la processualità dell'emergenza del fenomeno che, se seguiamo le ipotesi maggiori informate dalle teorie modulariste della percezione categoriale, erano fino ad ora rimasti silenti. Tuttavia, il tipo di spiegazione fornita, per stessa ammissione degli autori degli studi (Urgen et al. 2015, 2018), non riesce a rendere conto della specifica circostanza dell'UVE, in cui l'aspettativa disattesa riguarda l'incontro con il robot umanoide e crea questo specifico effetto straniante. Abbiamo ricondotto tale difficoltà alla posizione teorica del *predictive coding*, che si concentra esclusivamente su ciò che accade nella mente dell'individuo, confinata nei limiti incarnati del cervello umano, e al retaggio computazionalista ancora vivo in questo tipo di teorie. L'aspetto incarnato e pratico della percezione sottolineato dall'enattivismo ci ha consentito di avanzare di qualche passo nella considerazione di fattori ulteriori a quelli considerati essere al cuore del processo cognitivo nel *predictive coding*, permettendoci di vedere come sia necessaria la considerazione di elementi come la socialità, la corporeità, la contestualità dell'evento, la pragmatica dell'azione e l'espressività della pratica percettiva, tutti aspetti rimasti ai margini negli altri approcci, al massimo inclusi tramite rappresentazioni nella struttura funzionale dell'architettura cognitiva dell'agente, ma difficilmente considerati avere un ruolo attivo e costitutivo nella formazione dell'esperienza umana.

Nel nuovo quadro teorico che abbiamo costruito, quindi, entrano in gioco una serie di elementi di cui dare conto nel tentativo di fornire una spiegazione enattivista dell'UVE, che richiedono strumenti teorici e impianti epistemologici che possano tenere insieme la complessità dei fattori che abbiamo chiamato in causa. Arrivati a questo punto, abbiamo capito che il processo di riconoscimento del robot come somigliante all'umano ha un'importanza sostanziale per l'emersione dell'UVE, ma solamente se inserito all'interno del micromondo dell'individuo: vedere il robot come *humanlike* implica una serie di azioni che vengono enattivate proprio in concomitanza con la percezione dei suoi tratti estetici e acquistano un significato pragmatico alla luce delle possibilità di intervento reale o potenziale. Tuttavia, non abbiamo ancora mai specificato la natura di queste azioni, né quali siano le aspettative dell'essere umano in questo tipo di situazioni. Una serie di domande, quindi, rimane ancora irrisolta: cosa significa, in termini pragmatici, che l'essere umano percepisce il robot come somigliante all'umano? Quali sono le aspettative aperte dal micromondo della HRI con robot umanoidi? Come si presenta l'UVE in questo contesto?

Per risolvere questi dubbi, nel prossimo paragrafo ci affideremo, da una parte, alla formulazione enattivista del *predictive engagement*, i cui strumenti concettuali ci permetteranno di comprendere come il processo di disattesa delle aspettative non comprenda solamente le operazioni neurocognitive svolte dal cervello, ma l'intero apparato enattivo che lega indissolubilmente corpo e cervello dell'organismo con l'ambiente all'interno del quale è inserito e con la storia individuale dell'agente che enattiva significati tramite pratiche percettive incarnate; dall'altra, attraverso le descrizioni

dell'interazione sociale e intersoggettiva fornita dall'approccio del *Participatory Sense-Making*, potremo comprendere quali siano le aspettative nei confronti del robot nel momento in cui questo viene riconosciuto come un agente interazionale simile all'essere umano. Attraverso l'integrazione di queste due prospettive, sarà finalmente possibile fornire una descrizione dell'UVE non limitata a processi cognitivi interni o alla conformazione funzionale del cervello, ma in linea con una visione radicalmente *embodied* della mente e dell'esperienza umana.

### **5.3 Interazioni di cristallo: coordinazione, abiti e aspettative nella HRI**

#### *4.4.8 Predictive Engagement: abiti e aspettative in interazione*

Date le caratteristiche presentate dal *predictive coding*, questo sembra a prima vista non poter essere più distante da una teoria enattivista della mente e della cognizione (Kirchhoff 2015; Gallagher & Allen 2019; Di Paolo et al. 2021). Tuttavia, cerchiamo di non buttare via il bambino con l'acqua sporca: l'accento posto dalla dimensione predittiva della cognizione risulta altamente compatibile con una visione enattivista e incarnata della cognizione, tanto da costituire attualmente una delle tematiche più dibattute all'interno della *koinè* enattivista (Bruinenberg et al. 2018; Kirchhoff & Robertson 2018; Kirchhoff (Ed.) 2018; Gallagher & Allen 2018; Gallagher 2020b; Gärtner & Clowes 2017; Di Paolo et al. 2021; cfr. nota 180). Comprendere come sia possibile rivalutare la prospettiva predittiva sulla cognizione senza necessariamente fare appello a meccanismi cognitivi interni alla mente costituisce un passo fondamentale per arrivare a formulare un approccio all'UVE non internalista e che non riduca la spiegazione a meccanismi funzionali localizzati esclusivamente nell'architettura neurocognitiva dell'agente.

Come già brevemente anticipato, dunque, l'alternativa al modello computazionale e funzionalista è già stata formulata all'interno dell'approccio enattivista sotto l'etichetta di *predictive engagement* (Gallagher & Allen 2018; Bizzari & Hipolito 2016; cfr. Schlicht 2023). In sintesi, il *predictive engagement* è definito «a dynamical adjustment in which the brain, *as part of and along with the larger organism*, actively responds in ways that allow for the right kind of ongoing attunement with the environment – and environment that is physical but also social and cultural» (Gallagher & Allen 2018: 2634, corsivi degli autori). L'enfasi, in altre parole, non risiede nei processi cognitivi interni come nel modello funzionalista, ma nel modo in cui la rete di relazioni costituita da corpo, cervello e ambiente si assesta e si evolve in tempo reale sulla base delle interazioni incarnate che l'agente realizza ed esperisce. Infatti, il *predictive engagement* si caratterizza principalmente per la sua struttura distribuita nell'ambiente, le cui radici non sono rintracciabili solamente nelle modalità di funzionamento del cervello, ma nelle dinamiche di aggiustamento dell'agente nella sua attività di *sense-making*: il *predictive engagement* «it's doing a doing, an enactive adjustment, a worldly

engagement with anticipatory and corrective aspects already included. It is a loop that also navigates through the body and environment and form a whole» (Bizzari & Hipolito 2016: 95).

Possiamo cercare di riassumere le caratteristiche che definiscono il *predictive engagement* guardando alle differenze con l'approccio computazionalista del *predictive coding*:

- a) La prima differenza è relativa al ruolo che l'errore di predizione ricopre all'interno dei due modelli. Infatti, se per il *predictive coding* la costante minimizzazione dell'errore di predizione è l'attività principale svolta dal cervello, che cerca costantemente di appianare il divario tra ciò che si aspetta di percepire e ciò che realmente percepisce aggiornando i modelli previsionali o correggendo la modalità predittiva (Howhy 2013; Clark 2016), per il *predictive engagement* il *mismatch* si inserisce all'interno di un più ampio quadro interazionale. In realtà, per il *predictive engagement* non ha senso parlare di errore nei termini adottati dal *predictive coding*, dal momento che la struttura predittiva si configura in tutt'altra maniera. Il corrispettivo dell'errore, in questo caso, consiste nella dinamica di aggiustamento insita nel *sense-making*: «the constant bringing forth of signification [...] we may describe as a permanent lack in the living» (Varela 1991a: 8). La principale attività dell'organismo, in questo senso, è quella di assecondare alcune delle dinamiche del senso in cui si trova avviluppato, in un ciclo di attese e valutazioni che non hanno il carattere del costante monitoraggio dello stato di errore interno. L'errore è un aspetto del ciclo interazionale, il cui baricentro non è sempre nel cervello: la mente non è separata dal corpo, ma distribuita nella rete di relazioni e interazioni in cui il corpo, il cervello e l'ambiente si incontrano. La dinamicità di cui si caratterizza l'approccio enattivista, allora, ricalibra l'importanza che il cervello ricopre nel modello, vedendo i processi neurocognitivi come un aspetto di un quadro più ampio, in cui il *mismatch* predittivo si caratterizza come un passaggio obbligato nella precarietà del *sense-making*: non costituisce l'eccezione al buon funzionamento della cognizione; piuttosto, è dato dallo sfondo normativo pragmatico su cui il *sense-making* si esercita.

Enactivists suggest that the brain is not located at the center, conducting tests along the radiuses; it's on the circumference, one station amongst other stations involved in the loop that also navigates through the body and environment and forms the whole. Neural accomodation occurs via constant reciprocal interaction between the brain and body, and notions of adjustment and attunement can be cashed out in terms of physical dynamical processes that involve brain and body, including automatic and peripheral nervous systems. (Gallagher & Allen 2018: 2365)

- b) Un altro aspetto su cui i modelli si dividono riguarda il ruolo ricoperto dall'azione. Per i modelli computazionalisti, l'azione è uno degli strumenti attraverso cui l'organismo riduce l'errore di predizione, regolando l'interazione con il mondo affinché le predizioni risultino almeno minimamente confermate e l'errore minimizzato. Nel *predictive coding*, il movimento

agentivo risulta solamente propedeutico all'aggiustamento e all'allineamento delle predizioni e dell'evento percettivo, senza svolgere necessariamente la funzione di elemento costitutivo del processo, né è considerato parte imprescindibile nella struttura della cognizione, dato che spesso tutto quello che conta è già presente nella struttura funzionale interna che coincide con l'architettura cerebrale (Howhy 2013)<sup>183</sup>. Così, le azioni da svolgere vengono selezionate allo scopo di allineare le percezioni con i modelli predittivi (Hohwy 2020)<sup>184</sup>. Al contrario, come abbiamo visto, l'azione nella prospettiva enattivista costituisce un aspetto fondamentale della percezione e della cognizione, che ne realizza la stessa possibilità: è parte costitutiva del ciclo percettivo, un'attività incarnata in cui il significato emerge come l'aggregazione temporanea di elementi salienti all'interno del quadro interazionale, intrinseca parte della pratica percettiva dal momento che percezione è azione. Questo implica una ristrutturazione della comprensione del processo predittivo, che si dispiega nel tempo e nello spazio e che si configura come una rete di pratiche predittive, più che nei termini di una singola azione di inferenza atta a raccordare il momento:

[The predictive coding account] goes on to ignore the dynamic features of the environment, treating it as a set of states of affairs – as populated by objects having certain properties. [...] However, the world is not just comprised of states of affairs, it 'is dynamic, including as it does the dynamic activities of social others, the brain and body of a focal individual are not the only possible sources of dynamic organizing processes' (Hutchins 2011: 443). Such a perspective on 'the context' highlights that human everyday activity is embedded within various forms of patterned practices. (Kirchhoff 2015: 78)

Attraverso le parole di Hutchins, Kirchhoff richiama l'attenzione sul fatto che il *predictive coding* pensa al mondo come discretizzabile e, nella maggior parte dei casi, già discretizzato, un insieme di stati di cose che si mostra alla percezione dell'agente nella sua struttura atomizzata. Così, l'attività principale della cognizione consiste nel mettere insieme questi stati di cose, cercando di procedere nel modo più familiare possibile. La struttura dinamica dell'enattivismo, d'altra parte, mette da subito in discussione questa visione (Varela et al. 1991) e pensa al mondo come innervato da *affordance* che lo diagrammatizzano, ma sempre in relazione alle capacità interazionali e cognitive dell'organismo e non in termini assoluti: non esiste una oggettiva segmentazione del reale, come abbiamo visto nel caso della *humanlikeness*, ma una configurazione locale che si dà in accordo alle capacità filogeneticamente evolute dell'organismo e individualmente coniugate nella sua storia di interazioni passate che costituiscono le capacità attuali di esercizio del *sense-making*.

<sup>183</sup> «[...] normal perception is nevertheless at one remove from the real world it is representing» (Hohwy 2013: 137).

<sup>184</sup> «Action may be undertaken for *epistemic value*, to reduce uncertainty (e.g., by looking to confirm the cup's location), and for *utility*, to obtain desired states (e.g., drinking from the cup to quench thirst). In both cases, the rewarding state corresponds to prediction error minimization» (Hohwy 2020: 212, corsivi dell'autore).

L'azione, in questo caso, è l'elemento centrale, condizione fondante in quel loop dinamico con la cognizione che struttura le possibilità interazionali dell'agente<sup>185</sup>:

Action is not something happening in the brain, and it's not just providing new sensory input for neural processing; it's what the whole organism does in interactions with the environment, or under different description, what a person does in the world, and this changes the world as much as it changes the brain. (Gallagher 2023: 275)

- c) Questi primi due punti conducono a una delle differenze più significative tra i due modelli, che riguarda l'origine del significato percettivo. Nel *predictive coding*, il significato degli stimoli percettivi proviene dall'organizzazione dei modelli interni all'architettura neurocognitiva, i quali categorizzano l'input e lo interpretano secondo gli schemi predittivi disponibili costruiti sulla base di minimizzazioni degli errori avvenuti in precedenza e alle regole statistiche modellizzate secondo la teoria dell'inferenza bayesiana. Invece, in una formulazione decisamente differente, per il *predictive engagement* il significato percettivo è enattivato a partire dalla realizzazione pragmatica della pratica percettiva, non attraverso la rappresentazione di ciò che avviene all'esterno della mente, ma considerando l'intera rete di relazioni tra gli elementi costitutivi della cognizione *come mente*: il significato percettivo è il risultato generato da – e sempre re-inserito nel – processo di *sense-making* dell'organismo, il quale a sua volta informa lo svolgimento dell'interazione e ne subisce gli effetti. Il significato, quindi, passa dall'essere derivato dal modo in cui la rappresentazione viene valorizzata nei passaggi tra diversi meccanismi funzionali (statistici, inferenziali, bayesiani, percettivi, ecc.) a costituire l'esito in itinere dell'attività caratterizzante dell'organismo vivente, che si dispiega nelle occasioni concrete di realizzazione tramite l'azione e le pratiche percettive e che viene consumato nell'impresa di continuare a vivere e comprendere quali siano le possibilità di intervento permesse nell'ambiente di riferimento. La differenza maggiore, in questo caso, è che il significato non è solamente il risultato di un processo a valle di un'operazione rappresentazionale, ma pervade tutto il processo di percezione, comprensione e azione nell'ambiente: ogni momento è significativo per l'agente, e il significato si addensa nello stesso frangente in cui qualcosa viene percepito, ovvero inquadrato sotto un certo rispetto agentivo, come abbiamo già sottolineato in precedenza (cfr. §. 3.4.1; §4.4).

Meaningful encounters with the world imply a perceiving agent with some basic motivation to perceptually engage her surroundings. Schemata of sensorimotor contingencies give an agent the *how* of perception, a tacit knowledge of potential sensorimotor engagements, without giving its *why*, which depends on latent valences that

---

<sup>185</sup> Questo aspetto è chiaro, ad esempio, in Bruinenberg et al. (2018): «[...] it is *action* that does the work of actually minimizing surprisal. Actions change an organism's relation to the environment, thereby changing the sensory states of the organism [...]».

push or pull in one direction or another for attention and for potential sensory-motor engagement, reflecting, for example, a degree of desirability. (Gallagher 2023: 234)

- d) Infine, l'ultima differenza consiste nella struttura e nel contributo fornito dai *prior* nel processo di formazione delle aspettative. Nel *predictive coding*, i *prior* individuano le conoscenze precedenti, localizzate a un livello alto della gerarchia di processamento e comprendenti sia le credenze raccolte nel corso dell'esperienza individuale, sia il portato evolutivo dell'organismo biologico (Hohwy 2020). In questo quadro, i *prior* si costituiscono come equivalenti a «folk-psychological propositional beliefs» (Hohwy 2020: 216), che direzionano il processo cognitivo dall'alto fornendo la base per l'elaborazione delle predizioni dell'errore e il banco di prova per la soddisfazione delle aspettative. In linea con l'inserimento degli elementi euristici in una più ampia visione di insieme, non proposizionale e orientata all'azione, nel *predictive engagement* i *prior* acquistano una caratterizzazione ecologica, radicata nel corpo e strettamente legata alla storia sensorimotoria dell'organismo, delle sue disposizioni affettive e delle sue concrete realizzazioni in interazione: «[...] priors, which include affect, are not just in the brain, but involve whole body dispositions and adjustments – ‘anatomically informed priors’» (Gallagher & Allen 2018: 2636). Nella prospettiva enattivista, sono le esperienze precedentemente enattivate in contesti interazionali con l'ambiente e con altri individui a costituire la struttura dei *prior*, non delle rappresentazioni concettuali delle azioni, relative al singolo e puntuale processo di minimizzazione dell'errore di predizione, ma un pattern di conoscenze incarnate che instradano il percorso del senso e lo modellano, sempre in congiunzione alle azioni e alle pratiche percettive realizzate. In questo senso, un ruolo primario è ricoperto dal corpo, che costituisce, tra tutti, il *primo prior* (Allen & Tsakiris 2018), influenzando costitutivamente sulle modalità percettive e cognitive dell'organismo e determinando le possibilità di azione in ogni situazione. Ancora una volta, si ribadisce dunque il ruolo costitutivo dell'*embodiment*, sorgente delle azioni incarnate che rendono un organismo cognitivo tale.

Uno studio citato da Gallagher e Allen (2018: 2363) e da Kirchhoff (2015: 85) può aiutarci a chiarire il modo in cui un approccio come quello suggerito dal *predictive engagement* sia profondamente diverso dalle declinazioni computazionaliste del *predictive coding*<sup>186</sup> e riesca a rendere conto di fenomeni complessi che avvengono nella percezione senza ridurre il problema a meccanismi cognitivi interni, e anzi tenendo conto di tutto il portato interazionale ed esperienziale della percezione. In particolare, alcuni studiosi (Barrett & Bar 2009; Barrett & Simmons 2015; Chanes & Barrett 2020) hanno proposto l'ipotesi della predizione affettiva (*affective predictive*

---

<sup>186</sup> E anche rispetto ad alcune formulazioni internaliste del *predictive processing* (Clark 2016; cfr. Gallagher & Allen 2018: 2633-2634).

*hypothesis*) per dare conto dello stretto legame tra azione, struttura della percezione e significato attribuito all'oggetto intenzionato. Tale ipotesi «implies that responses signaling an object's salience, relevance or value do not occur as a separate step after the object is identified. Instead, affective responses support vision from the very moment that visual stimulation begins» (Barrett & Barr 2009: 1325). In altre parole, «people do not wait to evaluate an object for its personal significance until after they know what the object is. Rather, an affective reaction is one component of the prediction that helps a person see the object in the first place» (2009: 1331): «conscious percepts are indeed intrinsically infused with affective value, so that the affective salience or significance of an object is not computed after the fact» (2009: 1328). Da questo studio, ciò che risulta evidente è che il processo di percezione e predizione non avviene per step sequenziali in cui una rappresentazione passa da un livello funzionale all'altro per essere raffrontata a modelli proposizionali di *prior* percettivi, acquistando così significato come accade per il *predictive coding*, ma si configura come un processo di aggregazione immediata di significati che si realizzano nell'intreccio di diverse istanze costitutivamente rilevanti: l'*embodiment* e le capacità percettive dell'organismo a esso legate, la disposizione affettiva dell'agente, l'occasione concreta di realizzazione della pratica percettiva, le conoscenze incarnate che dirigono il riconoscimento degli *affordance* e la loro efficacia nell'insieme dinamico del *sense-making*.

[...] we do not just perceive the environment in a cold, detached fashion; rather, perceiving is already to be in an affective state such that bodily affective changes are part and parcel of the phenomenology of experience. [...] an organism's model when perceiving an object, person, and/or situation is already 'set up' – viz. it is configured – in accordance with prior associations involving not merely the internal circuitry but an entire body adjustment. (Kirchhoff 2015: 85).

La situazione si chiarisce ulteriormente se recuperiamo le caratteristiche del *micromondo* fornite da Varela (1992b) e rileggiamo le modalità attraverso cui l'organismo si situa nell'ambiente e caratterizza il suo *sense-making* tramite l'intreccio di pratiche percettive, formazione delle aspettative e *breakdown* dell'interazione. In cfr. 5.1.1, abbiamo identificato il micromondo attraverso alcune caratteristiche principali: in primo luogo, la sua ricorsività, tale per cui il micromondo tende a realizzarsi come panorama percettivo di possibilità di azione, costituendosi come lo sfondo regolare in cui il *sense-making* incarnato si realizza, caratterizzato dalla storia dell'organismo; in secondo luogo, abbiamo visto come il micromondo sia condizionato nella sua realizzazione dalla pragmatica dell'interazione, che direziona le salienze percettive e indirizza l'attenzione in relazione alle dinamiche concrete di sviluppo dell'azione. Il micromondo, in questo caso, inserisce il modello predittivo della cognizione nel quadro più ampio e familiare dell'interazione viva con il mondo, identificabile con quell'atmosfera che determina la grana dell'interazione. Considerare l'agente come sempre inserito all'interno di un micromondo che apre corsie privilegiate del significato tramite

l'enattivazione di *affordance* sociali e ambientali ci consente di concentrare lo sguardo sul funzionamento del processo di creazione di possibilità agentive in risposta a interazioni attese dall'agente.

Ora, la prima caratteristica resa intellegibile da questo connubio tra micromondi e *predictive engagement* consiste nel fatto che l'operazione di predizione si realizza distribuita nell'ambiente, e non confinata nella architettura funzionale del cervello. La nozione di micromondo concettualizzata da Varela ci permette di comprendere come la pratica percettiva, lungi dall'essere considerata l'operazione di mappatura e categorizzazione dello stato di cose del mondo esterno, sia sempre instradata da condizioni materiali di realizzazione, da contingenze agentive e ambientali immediatamente visibili all'agente come salienti nel processo di costruzione di senso. L'esempio portato dallo stesso Varela è emblematico in questo senso: quando mi accorgo di non avere più il mio portafogli dove *solitamente* lo ripongo, il micromondo che si realizza è costituito dalle possibilità agentive che posso mettere in pratica per ritrovare l'oggetto perduto, ed è lo stesso orizzonte percettivo che si modella sulla base delle operazioni rivolte al raggiungimento dell'obiettivo. In questo senso, i significati percettivi non sono generati dal recupero a livello alto di concetti e proposizioni, ma dall'immediata composizione di un piano di intervento in cui alcune possibilità risultano più adeguate di altre. Riletto attraverso gli strumenti del *predictive engagement*, il momento in cui mi accorgo di non avere più il portafogli con me costituisce un momento di necessaria ricalibrazione dei significati che, fino al momento immediatamente precedente, popolavano il micromondo all'interno di cui ero immersa, ma che ora vengono messi sullo sfondo per favorire l'emersione di un nuovo pattern di possibilità di azione ed elementi degni di attenzione. Non è isolando il processo di calibrazione dell'errore di predizione rispetto al percepito atteso, quindi, che comprendiamo come i significati si ristrutturano a fronte di un imprevisto, ma prendendo in considerazione l'equilibrio precario tra azione, percezione ed esperienza nell'intero quadro interazionale dell'agente, distribuito nelle *affordance* che costruiscono il micromondo corrente. L'atmosfera è cambiata, e con lei l'insieme delle potenziali pratiche percettive convocabili.

In questo quadro, le aspettative si configurano come elemento centrale della scena pratica, ma non più di natura concettuale e proposizionale, rappresentate all'interno della struttura funzionale come elemento astratto di innesco delle predizioni. Il secondo aspetto che è fondamentale sottolineare nella congiunzione di teoria dei micromondi e *predictive engagement*, allora, riguarda la differente considerazione delle aspettative e del modo in cui queste si caratterizzano. Le aspettative sono, innanzitutto, incarnate e legate alle possibilità di azione, delle *embodied expectations* che condizionano il modo in cui intraprendiamo le azioni e ne commisuriamo l'efficacia: «the visceral patterns of response [...] understood in terms of affect programmes which follow basic scripts» (Hutto



2006: 165). In questo senso, rileggere l'importanza dei *prior* nel processo di *predictive engagement* realizzato nel *sense-making* significa spogliarli della caratterizzazione astratta e proposizionale e, invece, vederli come le disposizioni all'azione enattivate dal processo percettivo che prevede il coinvolgimento dell'azione. In altre parole, ciò che accade è che il nostro corpo porta con sé una serie di prospettive sulla percezione, e le nostre possibilità agentive sono condizionate da una storia in cui il nostro stesso corpo è stato implicato in processi di comprensione attiva e che si sono sedimentati in *abiti percettivi sensorimotori* (Di Paolo et al. 2017: 142-146). Questi si configurano come schemi di azione non storici concettualmente, ma sedimentati nella stessa struttura sensorimotoria che lega corpo e ambiente, elementi costitutivi della capacità agentiva dell'organismo che acquistano una propria identità nell'insieme delle capacità esprimibili dall'agente. Le aspettative non vengono convocate tramite richiamo concettuale innescato da un riconoscimento categoriale, come suggerito del *predictive coding*, ma dagli intrecci pratici in cui si realizzano gli abiti sensorimotori enattivati dalla situazione interazionale. In sostanza, gli abiti *embodied* generano aspettative *embodied*, che si correlano nell'effettivo sviluppo del micromondo che stiamo abitando.

Ancora una volta, si pensi alla situazione in cui smarriamo il portafogli e sarà chiaro il motivo per cui è necessario includere gli abiti incarnati e le aspettative che ne derivano per descrivere ciò che accade. L'esperienza del corpo, infatti, è intrinsecamente coinvolta: le nostre mani, ad esempio, si muovono automaticamente verso tasche e borse in cui *solitamente* troviamo il nostro prezioso portafogli, in cui lo abbiamo precedentemente riposto, enattivando quegli schemi di abiti sensorimotori che si sono sviluppati e organizzati attraverso ripetute esperienze simili. Quello che facciamo quando cerchiamo il portafogli consiste esattamente in un processo di attuazione degli abiti e calibrazione delle aspettative che coinvolge il corpo e il suo rapporto abituale con l'oggetto: è così che prevediamo le possibilità che il nostro portafogli si trovi dove lo abbiamo sempre riposto, anche se poi ci ricordiamo di averlo lasciato sul bancone del bar.

Ecco, allora, tornare utile il concetto di predizione, in questo caso però completamente trasformato nel suo ruolo all'interno del modello predittivo: non ci sono più *prior* astratti interni alla mente, ma disposizioni all'azione la cui attivazione si realizza nel rapporto agentivo con il mondo; non ci sono più elementi di natura concettuale a monte dell'operazione predittiva, ma fattori convocati concretamente nella pratica percettiva con un duplice ruolo: da una parte, instradare lo sviluppo dell'interazione con l'ambiente, configurando strade privilegiate che abbiamo già percorso e per cui sappiamo cosa aspettarci; dall'altra, rendere conto dell'efficacia del processo attraverso il suo portato normativo pragmatico. In altre parole, le nostre aspettative incarnate individuano il pattern di relazioni agentive precedentemente realizzate, sedimentate come abiti incarnati che dirigono la pratica percettiva nello stesso momento in cui viene enattivata: le nostre modalità di cercare il portafogli sono

il risultato di una storia di interazioni già avvenute, enattivate nel momento presente attraverso il richiamo a una strategia pragmatica tramite cui ho già attraversato quel micromondo. Quello che succede dopo, e che segue alla consapevolezza di aver dimenticato l'oggetto da qualche parte, è la presentazione di un altro micromondo, che richiama altre pratiche percettive, altre capacità di interazione con l'ambiente, altri abiti incarnati e altre aspettative a questi connesse, fino a che la situazione non cambia nuovamente e si trasforma in un altro micromondo, anche questo in qualche maniera familiare perché costruito dai nostri abiti percettivi e sensorimotori, in un processo che si sovrappone con – e forse realizza – lo scorrere della nostra esistenza.

La ricorsività individuata da Varela come uno degli elementi caratteristici della teoria dei micromondi si configura proprio nella creazione di pattern di abiti sensorimotori richiamati, ma anche nelle aspettative che questi implicano. Considerare questo aspetto come nucleo del *predictive engagement* ci consente, allora, di rileggere sia il processo che ci porta a riconoscere il volto del robot come *humanlike*, sia il fenomeno dell'UVE sotto una luce differente e, in seguito, a instradare l'indagine verso la comprensione della natura delle aspettative e degli abiti convocati nel micromondo della HRI:

- In §5.2.3 abbiamo osservato come, seguendo il paradigma enattivista, riconoscere qualcosa come *humanlike* significa enattivare il micromondo relativo a esperienze simili vissute in precedenza. Ora, possiamo aggiungere che, non soltanto il processo di riconoscimento avviene realizzando pratiche percettive che immediatamente rendono significativo l'aspetto e il comportamento del robot, ma anche che questo riconoscimento porta con sé una serie di abiti sensorimotori e di aspettative incarnate. Il riconoscimento della somiglianza comporta la convocazione di abiti sensorimotori, percettivi e di aspettative incarnate riguardo la possibilità di metterli in pratica. Il robot risulta *humanlike* non solo perché presenta dei tratti estetici che possono essere comparati a quelli degli esseri umani (come un volto, due occhi, una bocca, ed eventualmente delle mani, delle gambe, un busto, ecc.), ma soprattutto perché, in interazione, si colloca al posto di un altro essere umano e occupa la posizione relazionale aperta dai nostri abiti sensorimotori: riempie un vuoto nella struttura relazionale del nostro *embodiment* che, appena colmato, richiama a sé la serie di abiti e aspettative realizzate solitamente nelle stesse occasioni.
- L'UVE, quindi, assume una forma complessa in cui le aspettative su cui si misura il fenomeno derivano dagli abiti incarnati di interazione sensorimotoria. A differenza della VEH, che inquadra l'UVE come il risultato di un errore di predizione non minimizzato, nel quale la mente non riesce a risolvere internamente la discrepanza tra ciò che si aspetta e ciò che percepisce, la visione enattivista mostra come l'UVE possa emergere da un'interazione corporea in cui

l'errore non è semplicemente concettuale, ma profondamente legato al rapporto che le nostre aspettative *embodied* intessono con gli abiti agentivi e percettivi nel quadro pragmatico dell'esperienza corrente, rapporto generato tramite la condensazione di significati contestuali, portati sociali e culturali, e pattern di abiti sensorimotori e aspettative abbinati al micromondo attuale. Riusciamo così a comprendere che il fenomeno dell'UVE è molto più di un errore di predizione e che deve essere spiegato con strumenti che vanno oltre e al di fuori dei meccanismi computazionali interni alla mente: l'UVE si manifesta come una disarmonia tra abiti percettivi incarnati e un'esperienza situata nel micromondo valutata sulla base delle aspettative radicate nel corpo. È solamente in questo modo che è possibile recuperare il significato dell'UVE come *breakdown* nell'interazione che avviene tra umani e robot: la rottura si inserisce all'interno di un flusso continuo di sviluppo di pratiche percettive e sensorimotorie, un fenomeno puntuale ma che ha conseguenze sullo sviluppo dinamico dell'interazione incarnata.

Prima di riuscire a fornire un'analisi completa dei processi che portano all'UVE, allora, risulta necessaria un'ulteriore operazione teorica, in cui rendiamo esplicita la natura degli abiti incarnati sensorimotori e delle aspettative aperte dal micromondo della HRI. In questo capitolo, come in parte avvenuto in altre sezioni di questa tesi, abbiamo raccolto una serie di elementi che ora si rivelano propedeutici al compimento di questa operazione.

Nella sezione precedente, abbiamo raccolto sotto l'etichetta di artefatto sincretico un aspetto del robot che ci è servito a illuminare il processo di *intentionality detection* in HRI, in cui l'intenzionalità agentiva del robot è percepita dall'umano in forme di predisposizione all'azione coordinata. In questo capitolo, abbiamo individuato l'analisi dei *breakdown* come processo di verifica da inserire nel più ampio quadro del gioco dell'interazione, ovvero come momento chiave per analizzare le modalità in cui l'incontro con l'artefatto si sviluppa in maniere inaspettate: l'UVE, allora, è stato individuato come il fenomeno perfetto da indagare, vista la sua centralità negli studi di HRI. All'interno dell'analisi dell'UVE come *breakdown* dell'interazione, la somiglianza all'umano è stata riletta sotto la lente della percezione come *perceptually guided action*, aprendo alla possibilità di considerare la *humanlikeness* non solo come proprietà intrinseca dell'aspetto dell'artefatto, ma anche come una qualità esibita in interazione che innesca dinamiche provenienti dal mondo della socialità intersoggettiva. Quando percepisco il volto del robot come simile a quello di molti altri umani che ho incontrato nella mia esperienza, percepisco anche le possibilità di azione che vengono aperte da quelle qualità estetiche: i suoi occhi comportano la possibilità che possa vedere; le sue mani presuppongono la capacità che possa toccare e interagire materialmente con oggetti dell'ambiente. Inoltre, gli

strumenti concettuali del *predictive engagement* ci hanno garantito la possibilità di inserire l'UVE nel più ampio quadro dell'esperienza percettiva umana, e inquadrare questo effetto nella interpolazione di abiti sensorimotori e aspettative interazionali che da questi derivano.

Ora, quello che dobbiamo fare è comprendere quale sia il pattern di abiti e aspettative che contraddistingue la HRI e che, nei momenti di interruzione e rottura, può portare al risultato dell'UVE come *breakdown*, ovvero intervallo tra due micromondi e origine del passaggio tra questi. Il quesito che rimane ancora irrisolto, allora, riguarda il nesso tra abiti enattivati, aspettative pragmatiche e reali possibilità di interazione nel preciso incontro con robot umanoidi: cosa significa, nell'interazione tra umani e robot, che le aspettative vengono disattese? Quali sono queste aspettative?

Nel capitolo §4, abbiamo concluso la nostra argomentazione sul ruolo dell'intenzionalità nella HRI constatando la possibilità che il robot riesca a esibire nel comportamento forme di intenzionalità sincretica, la cui percezione da parte dell'agente umano si manifesta come indistinguibile, a livello di coordinazione corporea, da quella che si realizza di fronte a partner di interazioni umani. In quella sede, abbiamo sottolineato il fatto che questo aspetto sia la chiave che consente di collocare l'indagine sulla HRI all'interno di un regime di socialità, ed è per questo che abbiamo potuto eleggere il robot a plausibile compagno nel gioco dell'interazione: come testimoniato da diversi studi (Sciutti et al. 2012, 2013a; Vignolo et al. 2016a, 2017; cfr. §4.3-§4.4), se il robot mostra qualità sintetiche di comportamento che riproducono forme biologiche di movimento, e in connubio a realizzazione mimetiche di somiglianza estetica con l'umano, allora la risposta corporea, situata esibita dall'essere umano non sarà diversa da quella che avviene *solitamente* in altre interazioni di tipo sociale, in cui il partner è però un altro essere umano. I presupposti che abbiamo raccolto ci indirizzano, dunque, verso uno sguardo alle dinamiche di socialità aperte dal riconoscimento del robot come simile all'umano, imbricato nelle dinamiche antropomorfe che contraddistinguono la percezione umana degli artefatti artificiali, soprattutto se mostrano qualità estetiche che lo incentivano. Abbiamo constatato, inoltre, che l'effetto sincretico raggiunto dall'artefatto robotico sia dovuto anche al ruolo costitutivo dell'interazione reale: nell'incontro con il robot, l'assetto interazionale in cui si muove costantemente l'organismo biologico umano si struttura a includere possibilità di mutua coordinazione e rende un agente, che intrinsecamente non sarebbe in grado di sviluppare alcuna forma di comprensione genuina, capace di mostrare intenzionalità proprio attraverso lo scambio sociale attivo e le dinamiche di reciproco riconoscimento che si verificano a livello corporeo. In quella sede, abbiamo raccolto questi elementi per formulare delle riflessioni sul concetto di intenzionalità originale e per rivalutare l'intenzionalità come prodotto dell'*engagement* dell'agente con il suo ambiente di riferimento. In questo momento, queste informazioni tornano utili per comprendere il tipo di abiti percettivi e

sensorimotori enattivati dall'incontro con un robot umanoide e per analizzare il tipo di aspettative convocate come sfondo normativo dell'interazione.

#### 4.4.9 *Participatory Sense-Making, o la struttura dell'interazione sociale*

Ora, se il tipo di risposta corporea sollecitata di fronte al robot si rivela indistinguibile da quella che avviene in presenza di un altro essere umano, allora il modello migliore per comprendere quali siano per l'essere umano gli abiti sensorimotori incarnati e le aspettative che da questi derivano è quello che descrive cosa avviene, a livello corporeo, nelle interazioni sociali con altri esseri umani.

Abbiamo già fornito una prima disamina di un approccio enattivista alla cognizione sociale in §3.3, convocato per rendere conto di modelli alternativi che potessero sostituire i principi della ToM confluiti negli studi di HRI con un modello basato sulla percezione diretta e sulla comprensione incarnata dell'altro in interazione, ovvero l'*Interaction Theory* (Gallagher 2005, 2020). Il richiamo a questa declinazione del pensiero enattivista nel dominio della cognizione sociale ci ha permesso di riconfigurare l'impostazione metodologica generale dell'indagine, evitando l'appello a meccanismi interni di matrice inferenziale e fornendo le possibilità di allargare lo sguardo includendo molteplici fattori nel novero degli aspetti cruciali da indagare per comprendere la HRI.

Tuttavia, la IT non sembra poter rispondere esaustivamente alle questioni che abbiamo delineato in questo capitolo e che sorgono nell'indagine di fenomeni come l'UVE, dal momento che focalizza l'attenzione, nell'analisi dell'interazione, in un momento successivo alla costituzione della stessa. Nell'IT, infatti, viene dato per scontato che nell'interazione gli individui siano «*always already coupled*» (Gallagher 2005: 81, corsivi dell'autore). Questo perché, nell'approccio elaborato da Gallagher, sebbene l'importanza dell'interazione per la costituzione della cognizione sociale sia più volte messa in primo piano (Gallagher 2020: cap. 5; De Jaegher et al. 2010), il nodo centrale non è sui meccanismi che permettono la costituzione dell'interazione, né sui fattori che strutturano la stessa, ma sui presupposti epistemologici, filosofici e cognitivi che permettono la riformulazione dello sguardo sulla cognizione sociale da un punto di vista enattivista (cfr. *infra* §3.3). Quello che invece spetta a noi, in questo momento, è un'indagine in cui si mira a comprendere i meccanismi di istituzione della interazione sociale come forma specifica di incontro con l'altro, ovvero come questo tipo di incontro si realizza e si stabilizza come terreno interazionale (De Jaegher 2006). Quello di cui abbiamo bisogno, in sostanza, è un focus sulle modalità di costruzione dell'interazione per come si realizza, nell'atto pratico costituito dall'incontro tra due agenti.

La nostra attenzione, allora, si sposta verso un altro modello della cognizione sociale elaborato all'interno dell'approccio enattivista. Sviluppandosi come una specificazione delle teorie della cognizione sociale elaborate nell'IT, la descrizione dei processi di interazione che si costituiscono

nell'incontro intersoggettivo è stata fornita dalle teorie del *Participatory Sense-Making* (PSM; De Jaegher e Di Paolo 2007, 2008; Di Paolo et al. 2010; 2018; Cuffari et al. 2015; Fuchs & De Jaegher 2009; Torrance & Froese 2011; De Jaegher 2006). In questo frangente, il PSM può aiutarci a comprendere quali siano le modalità attraverso cui, abitualmente, l'essere umano entra in interazione sociale e costruisce delle possibilità di comprensione con altri agenti all'interno del proprio ambiente, e può aiutarci a specificare quali siano gli abiti incarnati richiamati all'interno di micromondi che presentano la stessa forma.

Per comprendere come il modello del PSM concettualizza le dinamiche incarnate che avvengono in interazione intersoggettiva, la prima nozione da prendere in considerazione è proprio quella dell'*interazione sociale*. Nell'ottica del PSM, l'interazione è inquadrata come

the coupling between an agent and a specific aspect of the world: another agent. Interaction is the mutual interdependence (or bi-directional coupling) of the behaviors of two social agents. Precisely which behaviors of the agents are implicated in this process will depend on the specific interaction and the situation in which it takes place. (Di Paolo et al. 2010: 27)

A partire dalla concezione del *sense-making* come attività primaria dell'organismo cognitivo (Varela 1979; Varela et al. 1991; Di Paolo et al. 2010; Di Paolo 2005), l'interazione sociale si caratterizza come il processo di costruzione di senso in cui è coinvolto anche un altro agente. Questo permette di distinguere questo tipo di interazione da altri, che non implicano l'intreccio dei processi di costruzione di senso, anche se sono comunque individuabili come incontri tra più agenti in un ambiente: ad esempio, lo scambio di calore che avviene tra corpi di individui stipati in un autobus pieno è considerabile come un'interazione, dal momento che quello che avviene è uno scambio tramite organismi viventi, ma non di natura sociale.

L'interazione sociale, d'altra parte, si configura a partire dal fatto che i) «in the social domain you are not perceiving a 'what' but a 'who' – a subject, not an object» e ii) «this 'who' that you are perceiving is herself a perceiver, and moreover a perceiver of you. She participates in the very same interaction in which you perceive her» (McGann & De Jaegher 2009: 13). L'interazione sociale, dunque, è caratterizzata esattamente dal processo di intreccio del *sense-making* degli agenti coinvolti, entrambi capaci di realizzare forme autonome di produzione di senso, ovvero come

[...] the regulated coupling between at least two autonomous agents, where the regulation is aimed at aspects of the coupling itself so that it constitutes an emergent autonomous organization in the domain of relational dynamics, without destroying in the process the autonomy of the agents involved (though the latter's scope can be augmented or reduced). (De Jaegher & Di Paolo 2007: 493).

In questo senso, l'interazione sociale mostra due caratteristiche che la distinguono dagli altri tipi di interazione. In primo luogo, l'interazione che si forma nell'incontro di due agenti si autonomizza

a sua volta, chiudendosi operazionalmente, e acquista la capacità di indirizzare il *sense-making* individuale di ognuno degli agenti coinvolti. Si pensi, ad esempio, a quelle occasioni in cui stiamo semplicemente avendo una conversazione con il nostro compagno, ma improvvisamente iniziamo a discutere. Molto probabilmente, nessuno dei due aveva intenzione di litigare, ma è comunque successo: è l'interazione stessa, nel suo farsi, che ha prodotto le dinamiche che hanno portato ad aprire una discussione, senza bisogno che uno dei due agenti creasse individualmente e volontariamente le condizioni per discutere. In secondo luogo, affinché possa aversi un'interazione di tipo sociale è necessario che ognuno degli agenti coinvolti *conservi* la propria autonomia interazionale nel dispiegarsi dell'incontro, ovvero che la capacità di intervenire come agente autonomo nell'interazione non venga inficiata dallo stesso incontro con l'altro<sup>187</sup>. Un esempio in questo senso è costituito da quelle circostanze in cui l'etichetta, la formalità o le contingenze pratiche sono così vincolanti da non permettere libertà di azione: contesti altamente formali o coercitivi (De Jaegher et al. 2010: 443), ma anche incontri tra agenti il cui potere di intervento sull'interazione e sull'altro è molto sbilanciato. Ad esempio, si pensi a due colleghi che lavorano insieme su un progetto: ognuno di loro deve poter proporre idee, fare domande e contribuire attivamente, ma se uno dei due colleghi accettasse passivamente tutte le decisioni dell'altro senza esprimere la propria opinione o intervenire con suggerimenti, l'interazione perderebbe la sua natura collaborativa, e lo stesso avverrebbe nel caso in cui uno dei due esercitasse una pressione sull'altro, obbligandolo ad accettare le sue proposte. Uno dei due colleghi potrebbe essere timido o disinteressato, oppure l'altro è un suo superiore e non è molto incline ad accettare suggerimenti. In ogni caso, l'autonomia di uno dei due è limitata, tanto a volte da non risultare determinante nello sviluppo dell'interazione. È solo mantenendo entrambi la loro autonomia di espressione e azione che si possono co-creare soluzioni e prendere decisioni in modo congiunto.

Se queste due condizioni vengono rispettate, quello che ci troviamo ad affrontare è un'interazione di tipo particolare, in cui gli agenti coinvolti sono entrambi capaci di realizzare processi di *sense-making* che hanno l'altro agente come uno dei target attentivi e come collaboratore nel processo di realizzazione del senso. È così che si realizza la commistione di intenzioni e comportamenti, la cui struttura viene resa intelligibile guardando alle modalità in cui il processo di costruzione di senso di ognuno degli agenti coinvolti viene influenzato dall'altro e dalla stessa pratica interazionale messa in atto, arrivando anche a generare risultati che altrimenti non si sarebbero potuti avere individualmente. Il PSM, quindi, si definisce come «the coordination of intentional activity in interaction, whereby the

---

<sup>187</sup>«Thus, social interaction has two characteristics: (1) there is a coupling, which is regulated so as to generate and maintain an identity in the relational domain. Thus, the resulting relational dynamics are autonomous in the strict sense of precarious operational closure [...] and define events and processes as either internal or external to the interaction. And (2) the individuals involved are and remain autonomous as interactors». (De Jaegher & Di Paolo 2007: 493)

individual sense-making processes are affected and new domains of social sense-making can be generated that were not available to each individual on her own» (De Jaegher & Di Paolo 2007: 497): «it is sense-making *performed* socially *enacted* as a shared practice» (Di Paolo et al. 2018: 74).

La capacità di entrare in questo loop dinamico, in cui il *sense-making* individuale viene condizionato dallo svolgimento dell'interazione stessa e produce risultati più complessi della somma degli sforzi dei singoli, trova il suo esempio più emblematico nello sviluppo ontogenetico. Alla nascita, infatti, il bambino viene subito inserito all'interno di un ambiente già pregno di significati sociali e culturali, anche se non possiede le capacità di comprenderli o sfruttarli a pieno. Nel corso dello sviluppo, in interazione quasi costante con il *care-giver*, il bambino impara a riconoscere i segnali significativi, a sfruttarne il potenziale in modalità prima non possibili, a coordinarsi nella danza dell'interazione in modi sempre più complessi (Trevarthen 1979, 1998; Reddy 2008). Questo accade non solamente grazie allo sviluppo delle facoltà cognitive e percettive del bambino, ma soprattutto perché viene preso nella rete di relazioni significative dal punto di vista corporeo, in cui il *care-giver* include il bambino e organizza le sue capacità di intervento sul mondo fin dai primi momenti dopo la nascita. Il bambino, quindi, osservando il proprio *care-giver*, ma soprattutto partecipando alle interazioni che vengono allestite da questo e che prevedono un suo contributo sensorimotorio e corporeo, arriva a strutturare le proprie capacità di interazione con l'altro e a enattivare capacità di comprensione e riconoscimento delle situazioni interazionali: in altre parole, nel processo di sviluppo ontogenetico, l'incontro con l'altro contribuisce sostanzialmente a generare le capacità di rilevazione e enattivazione di micromondi. Ciò che accade è che il *sense-making* individuale del *care-giver* realizza una struttura interazionale in cui è previsto anche il contributo attivo del bambino, il quale si trova a ricoprire un ruolo che prevede capacità più complesse di quanto il *sense-making* che è capace di realizzare in quel momento permette. Così, attraverso progressi minimi ma significativi, il bambino impara a ricoprire il ruolo che gli è stato mostrato nelle interazioni, imparando a riconoscere la posizione da occupare e a costruire una storia di interazioni passate che guideranno quelle future<sup>188</sup>. Nella formulazione data dal PSM,

We have an embodied, lived and immediate experience of the other in social situations and this is so because we are naturally and in an embodied way coupled with others from infancy. This coupling is possible because of the inter-modal link between proprioception and visual perception that exists from birth and connects the body schema (a set of subpersonal sensorimotor processes that are dynamically involved in movements and posture maintenance) with the perception of others. [...] *Basically, we know others because of our own embodied experience, not because their bodies look like ours, but because we experience them through our own bodies.* We are not confronted with an object to dismantle

---

<sup>188</sup> Il PSM, centrale nell'apprendimento delle capacità interazionali corporee e sensorimotorie del bambino, è stato utilizzato per descrivere anche il processo di acquisizione della pratica linguistica, che si pone in linea di continuità con l'acquisizione di capacità incarnate di gestione del senso (Di Paolo et al. 2018; Cuffari et al. 2015).



in encountering another but with someone that we already relate to at a very basic, bodily level. (Di Paolo et al. 2010: 26, corsivi nostri)

In sostanza, riconosciamo l'altro come possibile partner in interazione data la storia delle nostre precedenti esperienze incarnate di incontro intersoggettivo; non perché le sue sembianze somigliano alle nostre, ma perché «the world is experienced as a field of possibilities because of the directedness of our body» (Di Paolo et al. 2018: 43), ovvero perché il nostro corpo e le capacità sensorimotorie che da questo derivano sono state sfruttate e affinate nel corso dello sviluppo a riconoscere le possibilità interazionali e di concatenamento agentivo con l'altro. Prima di tutto, quindi, riconoscere l'altro significa concepire accordi sensorimotori in interazione con lui.

Arrivati a questo punto, è chiaro che le possibilità di interazione sono al centro della capacità di cognizione sociale dell'individuo e che costituiscono il nodo centrale da analizzare per comprendere come si struttura l'incontro con l'altro. Ma quali sono le caratteristiche esibite da questa capacità interazionale? Dal momento che l'interazione sociale è un processo esteso nel tempo e realizzato concretamente nell'incontro interindividuale, la necessità è quella di oltrepassare l'apparenza che la sua organizzazione sia realizzata sommando il singolo apporto di ognuno degli agenti coinvolti e analizzare, invece, il modo in cui si costituisce l'aggregazione di senso che organizza l'interazione nel suo complesso come sistema autonomo all'interno di cui il *sense-making* individuale si muove<sup>189</sup>. Il PSM individua la nozione di *coordinazione* come strumento teorico capace di rendere intelligibile la connessione tra gli aspetti temporali dell'interazione e le loro conseguenze per il *sense-making* congiunto e individuale (De Jaegher & Di Paolo 2007, 2008, 2009; Di Paolo et al. 2018). In particolare, la coordinazione è definita

[the] non-accidental correlation between the behaviors of two or more systems that are in sustained coupling, or have been coupled in the past, or have been coupled to another, common, system. A *correlation* is a coherence in the behavior of two or more systems over and above what is expected, given what those systems are capable of doing. (De Jaegher & Di Paolo 2007: 490)

Nel dominio dell'interazione sociale la coordinazione si dà in diversi modi. Quello più conosciuto è senza dubbio la sincronizzazione, come avviene nei casi in cui, camminando di fianco a una persona che conosciamo, il nostro passo si sincronizza e procediamo alla stessa velocità. Ma non solo: con coordinazione si intendono anche tutti quei fenomeni di *mirroring* (ad esempio, quando si è immersi in una conversazione e le espressioni facciali di chi parla vengono riprodotte, spesso involontariamente, dall'interlocutore, come meccanismo di *attunement* interazionale; Iacoboni 2009; Navarretta 2016), anticipazione (come avviene nel caso dello sguardo anticipatorio, cfr. §4.4; Falck-

---

<sup>189</sup> «We have to move beyond a view that defines interaction as simply spatio-temporal coincidence of two agents that influence each other. We must move towards an understanding of how their history of coordination demarcates the interaction as an identifiable pattern with its own internal structure, and its own role to play in the process of understanding each other and the world». (De Jaegher & Di Paolo 2007: 492)

Ytter et al. 2006; Riechelmann et al. 2021) e imitazione (ad esempio, un bambino che imita i comportamenti e i gesti del *care-giver* nella manipolazione di un oggetto, Thelen et al. 1980), ovvero tutte le forme di regolazione e interdipendenza che provengono dall'incontro pragmatico tra agenti incarnati. In questo contesto, quello che avviene è che l'agentività sensorimotoria del singolo individuo si aggancia a elementi delle azioni dell'altro, individuando punti di contatto e opportunità di concatenamento nella realizzazione della costruzione di senso comune e costruendo un piano di accordo su cui il contributo di ognuno viene a realizzarsi. La coordinazione ci aiuta a inquadrare l'interazione come un fenomeno progressivo e dinamico, che si sviluppa nello spazio e nel tempo, con il contributo necessario di entrambi gli agenti coinvolti ma come risultato non riducibile ad alcuno dei due singolarmente. In altre parole, la coordinazione è il risultato dell'interazione, e quindi una sua qualità emergente.

When there are coordination structures that help sustain the social encounter, and the encounter itself promotes coordination, this double link between encounter and coordination makes the collective pattern into an autonomous system according to our definition. This permits the identification of a specific interaction on the basis of the organization of its collective dynamics. (De Jaegher & Di Paolo 2009: 4)

Un importante aspetto, che consente alla coordinazione di emergere come risultato di un *attunement* dei *sense-making* individuali, viene individuato nel *ritmo* dell'interazione. In questo senso, il ritmo indica le diverse scale temporali su cui l'interazione può realizzarsi: individuarne la struttura significa riconoscere le possibilità di concatenamento adeguato allo svolgimento dell'interazione, non semplicemente in accordo alle intenzioni o al comportamento dell'altro agente, ma alla stessa forma dell'interazione che si sviluppa in modi legati al contesto, agli obiettivi e a fattori sociali e culturali. L'individuazione del ritmo dell'interazione è una capacità che si realizza a livello corporeo come identificazione di un pattern di azioni e pause su cui possiamo costruire le nostre possibilità di intervento e contribuire attivamente alla formazione dell'unità interazionale che abitiamo con l'altro. In altre parole, l'abilità di individuare il ritmo dell'interazione è

a capacity that is not strictly of an individual, but one that comes about in interaction and is changes by both the interactional processes and the individuals involved. We define this central capacity of social cognition as: flexibly temporally coordinating through the interaction with another person. Through such flexible coordination, the rhythm of an interaction can be adapted to changing circumstances, changes of goals, mood, etc. This capacity is crucially dependent both on the individual interactors and at the same time on the process of engagement that ensues between them in every interaction. A corollary of this is that the rhythmic capacity of a person is never complete outside of an ongoing interaction. (Di Paolo et al. 2010: 33)

Possiamo, quindi, affermare che i vantaggi mostrati dal PSM come modello di analisi dell'interazione sociale consistono: i) nella possibilità di analizzare la struttura agentiva interna dell'interazione, ovvero guardare sia ai contributi dei singoli nella costruzione dell'interazione, sia al

modo in cui l'entità interazionale autonoma condiziona e indirizza il contributo individuale; ii) nel considerare il contesto pragmatico della realizzazione dell'incontro, attraverso la struttura sensorimotoria e incarnata dei comportamenti e il ruolo della dimensione temporale nel dispiegamento della circostanza. A questi due aspetti, insiti nel PSM, se ne aggiunge un altro, che ci consente di adottare il PSM come modello per descrivere l'esperienza di fronte al robot.

Questo aspetto consiste nel iii) modo in cui vengono concepiti i momenti di *breakdown* dell'interazione sociale. Infatti, nell'ottica del PSM, che vede l'interazione dispiegarsi in diverse fasi e su più scale temporali, le rotture possono verificarsi come «obstacles and barriers to our actions, hesitations about what to do next, or maybe a general feeling of things not going smoothly» (Di Paolo et al. 2018: 43). A fronte di un impedimento, quindi, un ostacolo riconosciuto nello svolgimento progressivo del *sense-making*, si prospettano diverse forme di recupero, che si dispiegano sui molteplici piani in cui l'interazione si sviluppa. Abbiamo già visto nell'analisi dei *breakdown* elaborate da Varela (§5.1.1) che la presentazione di una rottura nel micromondo corrente apre a momenti di instabilità percettiva ed esperienziale che si concretizza nel passaggio a micromondi alternativi. In questo senso, anche nel PSM il *breakdown* apre l'interazione a sviluppi fino a quel momento inattesi, ma comprende anche l'altro agente nel processo di recupero dell'interazione. Anche nell'interazione sociale il momento di *breakdown* costituisce un elemento importante del processo interazionale, e il PSM sottolinea il ruolo che questo tipo di esperienze hanno nel rafforzamento dell'interazione, in quanto momento di conoscenza reciproca e occasione di disvelamento delle dinamiche interazionali:

Sustained interactions can be expected to have undergone several instances of loss and regain of coordinating structures, each of them leaving the interactors slightly better able to remain in such interaction or reinitiate it – experientially, we often perceive some interactions as improving over time, and the recovery from a breakdown as a sort of learning in which our previous moves acquire new contextual significance. Viewing interactions from this perspective is akin to understanding the growth of an adaptive system. Such a view allows us to connect interaction dynamics with sense-making. (De Jaegher & Di Paolo 2007: 496)

In sostanza, quello che accade è che, alla percezione di una rottura, la costruzione di senso partecipata distribuisce possibilità di intervento in direzioni che possono riuscire a recuperare l'interazione, in cui gli agenti collaborano al fine di sostenere il corso degli eventi, anche se inaspettato: «the recovery from such coordination breakdowns involves a mutual accommodation between the participants that, if successful, moves the interaction forward into a new state of coordination» (Di Paolo et al. 2018: 74). Se riprendiamo l'esempio della stretta di mano mancata che abbiamo già sviluppato, possiamo comprendere meglio come l'aggiustamento conseguente al *breakdown* possa risolversi sia con il nostro intervento, in cui cerchiamo di toglierci dall'imbarazzo magari facendo finta di stare afferrando qualcosa, ma anche plausibilmente facendo affidamento sulle

capacità interazionali del nostro interlocutore: probabilmente, cercheremo di mostrare la mano in maniera più esplicita, spostandoci di fronte a lui per manifestare la nostra presenza o rientrando nella sua visuale, in questo modo rendendo palese l'*affordance* sociale che gli stiamo proponendo, e così cercare di recuperare l'interazione insieme. Il nostro interlocutore potrebbe aver dapprima esitato nella stretta, ma alla sollecitazione sensorimotoria esercitata tramite esibizione del gesto potrebbe voler rispondere intraprendendo quella pratica interazionale che è la stretta di mano. I *breakdown*, d'altra parte, possono anche aiutarci a comprendere meglio le dinamiche di interazione che possono svilupparsi: se il nostro interlocutore, anche dopo i nostri tentativi di recupero, non ci stringe la mano, allora capiremo che quel gesto non è accettato, per motivi legati forse alla sua provenienza culturale, ad una timidezza estrema o magari a un brutto carattere. Quello che è importante è che avremo imparato qualcosa da quella rottura interazionale, che molto probabilmente instraderà situazioni simili e, sicuramente, gli incontri futuri con la stessa persona.

Failures in understanding an other's behavior are not exceptional. On the contrary, they form part and parcel of the ongoing process of social understanding. More even, misunderstandings are the pivots around which they really interesting stuff of social understanding revolves. In these instances where coordination is lost, we have the potential to gain a lot of understanding. [...] Loss and recovery of coordination are mediated through the interaction itself – a coordination which, as we have seen above, is not a constant attunement, but rather a variable, sometimes even discontinuous process. It is when the flow of movement across me and my interaction partner grinds to a halt that my or our sense-making activity can and often does change direction. When the ongoing flow of interacting and coordinating breaks, opportunities for redirecting sense-making open up. (De Jaegher 2009: 540).

Il PSM, quindi, si caratterizza come uno sguardo all'interazione sociale che non solo raccoglie i presupposti epistemologici di una teoria incarnata della cognizione, ma che sviluppa riflessioni in merito alla costituzione interna dell'incontro situato con l'altro. Riassumendo, gli aspetti illuminati dal PSM riguardano:

- i) L'interazione come entità autonoma, qualcosa che viene prodotta dall'incontro degli agenti ma che, allo stesso tempo, abbraccia i singoli sotto un'unica unità dialogante. È nella costruzione di questa unità, abbiamo visto, che il bambino impara a svolgere azioni complesse nel suo ambiente e in coordinazione con altri agenti, ed è così che si sviluppano significati incarnati condivisi consumati nelle occasioni di socialità.
- ii) La coordinazione come cifra distintiva dell'interazione sociale, la quale ci consente di osservare come i meccanismi di *attunement* che si verificano abbiano una struttura complessa, che si compone su diverse scale temporali e si costituisce nell'intreccio pragmatico delle azioni sensorimotorie attraverso la rilevazione corporea di un ritmo a cui è possibile agganciarsi, nei molti modi in cui la coordinazione può darsi.

- iii) L'inclusione dei *breakdown* come momento di sviluppo di una comprensione più profonda delle dinamiche e dei presupposti dell'interazione, che possono svelare occasioni di coordinazione inaspettata e che hanno la capacità di arricchire la comprensione individuale degli eventi e, quindi, consentire ulteriori opportunità di concatenamento.

Ora, a partire da questi principi, possiamo tornare a indagare la forma degli abiti sensorimotori e delle aspettative da essi generati per comprendere cosa, in termini corporei e situati, l'essere umano si aspetta di affrontare quando incontra qualcosa che occupa lo stesso posto di un essere umano.

In primo luogo, abbiamo visto come gli abiti sensorimotori siano generati dalla storia delle esperienze passate dell'agente, raccolti sotto la forma di pattern d'azione aderenti alla capacità agentiva dell'organismo, che vengono richiamati ogni qual volta la conformazione assunta dall'ambiente – il micromondo enattivato – si presenta percettivamente adeguato. In sostanza, nell'incontro con un altro essere umano, ciò che accade è che il micromondo che si apre si configura come una serie di possibilità di coordinazione, che riusciamo a intravedere proprio perché insite nel nostro modo di costruire il senso e acquisite in un percorso di enattivazioni lungo il corso della nostra vita. Per sapere quali sono le azioni adeguate in un determinato micromondo, possiamo fare affidamento sulle occasioni in cui i nostri comportamenti hanno già funzionato, in cui abbiamo appreso come muoverci e come coordinare le nostre azioni a quelle degli altri. Allo stesso tempo, il richiamo di abiti sensorimotori, con una propria struttura di coordinazione che prevede l'intervento dell'altro, comporta anche una serie di aspettative circa le modalità attraverso cui l'altro occuperà la posizione interazionale aperta dalla pratica abituale.

È in questo stretto legame tra abiti e aspettative che si configura la possibilità di comprensione reciproca e di svolgimento dell'interazione: solamente comprendendo le prospettive interazionali che l'altro ha aperto per me attraverso i suoi abiti incarnati, potrò concatenare la mia agentività nella produzione di senso condivisa, e lo stesso vale per lui; solamente individuando il pattern ritmico che contraddistingue il nostro incontro potrò anticipare le sue aspettative e occupare il posto riservato a me, e viceversa. La danza dell'interazione implica la capacità di esercitare la propria agentività quando permesso dalla interazione stessa, senza per questo inficiare le possibilità di accordo con le azioni del mio interlocutore, e anzi assecondarle per arrivare alla creazione condivisa del senso. Le norme sociali e culturali, in questi casi, ricoprono un ruolo fondamentale, dal momento che incasellano gli abiti sensorimotori attraverso cui è possibile portare avanti l'interazione e stabiliscono uno sfondo normativo – più o meno rigido in dipendenza dal contesto e dall'apparato culturale di riferimento – su cui l'efficacia delle azioni viene valutata.

Cosa implica, dunque, questo portato esperienziale, di abiti e aspettative, nell'incontro con il robot? I modelli che abbiamo esposto ci suggeriscono una lettura precisa. Quando si *percepisce* il robot come agente sociale (mimetico e sintetico: ovvero, sincretico) al pari dell'essere umano, gli abiti che si enattivano e le aspettative che si creano prevedono la possibilità di coordinazione tipiche delle interazioni avute con altri umani: automaticamente, nello stesso momento in cui percepisco il robot come *humanlike*, il mio corpo richiama l'assetto interazionale che contraddistingue il micromondo intersoggettivo, con il suo bagaglio di abiti sensorimotori pragmatici e aspettative circa la possibilità che il robot ricopra le posizioni lasciate vuote per l'altro, che l'altro deve riempire in accordo alle contingenze sociali specificate dal contesto e dall'interazione stessa. Tra queste aspettative, una in particolare ci pare rilevante da sottolineare: l'insieme delle attese aperte dagli abiti sensorimotori intersoggettivi comprende anche che l'altro, con cui si costituisce l'interazione, sia soggetto allo stesso tipo di processo e capace di sfruttare le condizioni poste dall'interazione e dall'interlocutore.

La concettualizzazione dell'interazione sociale da parte del PSM, infatti, ci permette di comprendere che il processo di costruzione dell'interazione sociale non accade in modo unidirezionale, ma che l'interazione prospetta l'aspettativa sensorimotoria dipendente dall'abito della coordinazione interazionale, che prevede che anche l'altro possa fare lo stesso. In altre parole, l'assetto dell'interazione sociale prevede che entrambi gli agenti implicati nell'interazione abbiano le stesse capacità di intervento sulla costruzione del significato, o che perlomeno queste siano compatibili, che entrambi siano capaci di comprendere il ritmo dell'interazione e che possano convogliare il proprio *sense-making* nell'operazione congiunta di realizzare il senso condiviso.

Si sarà intuito che, nel caso in cui l'interazione si verifichi con un robot, le circostanze non rispettano i criteri posti come necessari per un'interazione sociale. Al momento dell'incontro con il robot, infatti, il micromondo che si apre riversa all'interno dell'interazione tutto il portato storico dell'interazione tra umani, convoglia tutte le attese probabili in un incontro tra umani, impostando una serie di aspettative che provengono dalle precedenti esperienze intersoggettive e che hanno modellato la capacità di interagire fino a quel momento. Una delle conseguenze di questa bidirezionalità degli effetti dell'interazione, come abbiamo visto nel caso dell'intenzionalità, consiste nel fatto che il robot acquista possibilità interazionali proprio perché viene implicato in questo sguardo, in cui l'interazione conferisce a entrambi gli agenti lo statuto di interattanti: per l'essere umano, si tratta dell'ennesima occorrenza in cui si trova a dover ricoprire posizioni relazionali e a poter esercitare i suoi abiti sensorimotori in accoppiamento con altri agenti; per il robot, questo significa dotarsi di quel sincretismo che produce l'armonia degli effetti, realizzato grazie alle proprietà dell'interazione sociale istituite nell'incontro con l'umano.

Tuttavia, gli effetti che l'interazione ha sul robot non riguardano solamente ciò che l'artefatto acquista in termini di capacità percepita. Dal momento che l'interazione deve essere vista nel suo sviluppo spaziale e temporale, quello che si apre è un terreno in cui viene messa alla prova, sul piano interazionale, la capacità del robot di ricoprire le posizioni aperte dall'interazione, mostrando la capacità di concatenarsi e di coordinarsi alle azioni dell'essere umano seguendo il ritmo che pulsa nella reciprocità dell'azione: in sostanza, deve saper giocare al gioco dell'interazione. In un passo significativo, Hanne De Jaegher ed Ezequiel Di Paolo espongono la dinamica in cui gli agenti diventano interattanti proprio a causa della struttura dell'interazione sociale, aggiungendo anche una specifica che può aiutare a comprendere cosa accada nell'interazione con i robot:

Individuals co-emerge *as interactors* contemporaneously with the interaction. Considering how individuals are affected by the encounter leads to an additional requirement for defining the interaction as social: individuals as interactors must not lose their own autonomy in the process (even though the encounter may enlarge or diminish the scope of individual autonomy). This is a constitutive constraint necessary for defining the social. In its absence, if the autonomy of an 'interactor' were destroyed, the whole process would reduce to the cognitive engagement of the other agent with his world. The 'other' would become a tool, an object, or a problem for his individual cognition, making the engagement indistinct from non-social ones. (De Jaegher & Di Paolo 2009: 4)

In sintesi, uno dei presupposti affinché l'interazione sociale possa realizzarsi consiste nel fatto che entrambi gli agenti devono conservare la propria autonomia, pena la disgregazione dell'interazione e il cambiamento radicale del regime interazionale, che passa dall'essere sociale e intersoggettivo a un tipo di interazione meno complesso, in cui non si richiede la necessità di concatenare diversi *sense-making*. Questo aspetto viene ribadito ulteriormente in altri articoli sul PSM, specificando che «interactions are social as long as the autonomy of the agents is not dissolved. If one agent becomes the sole regulator of the coupling, as in the use of a tool, this is no longer social interaction» (De Jaegher et al. 2010: 443)<sup>190</sup>.

Eppure, sappiamo dalle indagini che ci hanno condotto fino a qui che il robot non possiede alcun tipo di autonomia genuina. Come accade per l'intenzionalità, il robot acquista la possibilità di essere considerato un agente autonomo grazie alla combinazione delle sue componenti sintetiche e mimetiche sotto l'aspetto sincretico, che si manifesta proprio perché l'umano allestisce le possibilità interazionali per il robot allo stesso modo in cui farebbe per un altro essere umano. Si potrebbe dunque pensare che, al fine di rendere conto del fenomeno dell'UVE negli HRI, si possa immaginare che sia questo il motivo dell'emersione dell'*uncanniness*: l'essere umano scopre che il robot non ha

---

<sup>190</sup> Questa visione in cui l'interazione con l'oggetto e la materialità viene vista come guidata esclusivamente dalle intenzioni dell'essere umano è stata messa in discussione proprio all'interno di un approccio enattivista alla cognizione materiale (Malafouris 2013). Non ci troviamo a convenire, dunque, sul fatto che l'interazione materiale sia realizzata senza il contributo attivo della materialità, concezione che potrebbe costituire anche una linea di ricerca interessante nel panorama degli HRI. Proveremo a sviluppare alcune riflessioni in merito nelle conclusioni al presente lavoro.

un'autonomia genuina e quindi lo assale un senso di inquietudine e disagio che si manifesta in una rottura incolmabile. Una spiegazione del genere, secondo noi, rimane comunque insufficiente nel rendere conto del fenomeno dell'UVE, dal momento che non considera due fattori:

- il fatto che, in altre occasioni, tali meccanismi non portano a effetti simili a quelli manifestati nell'UVE. Infatti, come sottolineato anche dagli stessi autori, occasioni in cui si manifesta la disgregazione dell'autonomia in uno degli agenti coinvolti non sono rare. L'autonomia interazionale dell'agente viene intaccata ogni volta che questa non può essere esercitata nella costruzione del senso condiviso: casi di coercizione, in cui un individuo costringe qualcuno a compiere un'azione, o casi in cui l'individuo non può esercitare la propria autonomia in interazione, come nel caso del sottoposto che lavora a un progetto con il suo superiore, sono esempi di sfaldatura delle capacità autonome dell'agente, ma non provocano in noi l'effetto di straniamento tipico dell'UVE. Potremmo essere turbati nel vedere una scena di questo tipo, come ad esempio succede guardando alcuni film, ma questo non è paragonabile alla sensazione descritta come il cuore dell'UVE.
- D'altra parte, i robot spesso riescono a performare azioni che sintetizzano alcuni tipi di autonomia organizzazionale: si pensi agli esperimenti sull'implementazione del movimento biologico esposti in §4.4, ma anche alle circostanze in cui è il robot, ad esempio, a iniziare una conversazione o a proporre di compiere un'azione. In questi casi, non si potrebbe parlare di una sfaldatura della capacità autonoma in interazione del robot, dal momento che quello che esprime è un modello sintetizzato di quella che viene realizzata dagli umani, che abbiamo visto mostrarsi nell'aspetto sincretico e costruirlo proprio in interazione.

Non ci sembra adeguato, quindi, ricondurre la rottura provocata dall'UVE a questo aspetto, ovvero al fatto che intrinsecamente il robot non può operare alcun tipo di intervento autonomo nell'interazione. Questo perché la capacità di autonomia, anche se in diversi gradi di presentazione, rimane comunque percepibile, salvo i casi in cui si esaurisce la batteria.

D'altra parte, potremmo pensare che l'UVE sia il risultato di un *breakdown* di altro tipo nell'interazione, qualcosa che si presenta come insolubile agli occhi dell'interattante umano. Questo tipo di ipotesi sembra riportarci ad alcune delle spiegazioni che abbiamo visto in §5.2.1, che prevedevano l'impossibilità di sorpassare lo scoglio cognitivo posto dal robot umanoide come scaturigine dell'UVE. Anche questo suggerimento, tuttavia, sembra incompatibile con la prospettiva adottata sia dal PSM, sia dal *predictive engagement*, sia dalle riflessioni fornite da Varela sui micromondi. Il *breakdown*, in queste epistemologie, viene incluso come fattore fondante dell'interazione con il mondo e, nel caso del PSM, dell'interazione sociale con altri sistemi.



Given the vulnerability of interactions and of individual sense-makers, breakdowns are unavoidable. But this is precisely what drives participatory sense-making. Without breakdowns, without the constant risk of misunderstanding or worse, there would be no participatory sense-making (i.e. no social understanding) at all. [...] When sense-makers engage in social interaction, it is the intertwining vulnerabilities of interaction and participants that make possible the ongoing various transformations of sense-making. [...] Breakdowns are not mere obstacles to social understanding that must be overcome, but lie at its root, even when interactions go smoothly. (Di Paolo et al. 2018: 75)

Senza *breakdown*, non c'è interazione sociale, sviluppo del *sense-making*, necessità di meccanismi di aggiustamento predittivo o costruzione di senso condiviso: il *breakdown* è parte costitutiva dello sviluppo dell'*engagement* con il mondo, sia che questo si dia all'individuo come rapporto con la sola materialità dell'ambiente, sia che si configuri in pratiche incarnate di percezione sociale. Motore della produzione di senso che mette in evidenza «[the] permanent lack in the living» (Varela 1992a: 8), la rottura è ciò che consente l'evoluzione, il passaggio di senso, la generazione di significato e lo sviluppo dell'interazione con altri agenti.

C'è qualcos'altro allora, che deve essere considerato, che si situa temporalmente in concomitanza con la messa alla prova dell'agente robotico nel campo dell'interazione. Questo perché anche i proponenti del PSM ammettono che «interaction with complex, but not fully autonomous systems (such as virtual characters or some social robots), can provoke an experience of engagement» (De Jaegher et al. 2010: 442), ovvero la percezione di trovarsi in presenza di un agente capace di rispondere all'interazione e di occupare quelle posizioni aperte dalla socialità. Quello che ci troviamo ad evidenziare, allora, è un complesso sistema di equilibri e bilanciamenti, aperti dalla pratica sociale e che hanno come effetto l'esperienza di una possibilità interazionale, i cui esiti possono portare a un'esperienza di tipo particolare:

In other words, if, over time, the kind of practices that individuals find themselves participating in comes to exhibit regularities – patterns of expectancy – then one might reasonably assume that disruptions in such patterned practices may result in a particular – non-anticipated – phenomenology. (Kirchhoff 2015: 86)

Se non possiamo rintracciare l'origine dell'UVE nel momento stesso della costituzione dell'interazione, allora, quello che proponiamo è provare a collocare l'UVE lungo il processo di sviluppo dell'interazione sociale, piuttosto che nel momento della sua strutturazione iniziale. Come suggerito da Winograd e Flores, infatti, vogliamo provare a indagare il modo in cui il *breakdown* in interazione si genera e i modi in cui questo prova ad essere recuperato.

[...] a person is not permanently trapped in the same kind of blindness. We have the potential to respond to breakdown with a shift of domains in which we enter into new commitments. Understanding is not a fixed relationship between a representation and the things represented, but is a commitment to carry out a dialogue within the full horizon of both speaker and hearer in a way that permits new distinctions to emerge. (Winograd & Flores 1986: 124)

Seguendo Varela, possiamo ipotizzare che sia proprio nel passaggio tra micromondi suggerito dal *breakdown* in interazione che si radichi l'effetto perturbante.

#### 4.4.10 L'UVE alla prova del gioco dell'interazione: la rottura dell'aspetto sincretico

Per comprendere in quale momento del processo interazionale si manifesta l'UVE come fenomeno di *breakdown*, riprenderemo le condizioni poste dal gioco dell'interazione che abbiamo sviluppato all'inizio di questo capitolo (§5.1). Abbiamo definito il gioco dell'interazione come una variante situata e basata sul ruolo del corpo del gioco dell'imitazione proposto da Alan Turing, sviluppato come banco di prova per rispondere alla domanda: *can robots interact?* All'inizio del capitolo, abbiamo delineato le differenze e le somiglianze con la versione originale del gioco, sottolineando l'importanza che la rottura dell'interazione acquista in ognuno degli assetti. A seguito delle riflessioni che abbiamo compiuto, possiamo elaborare ulteriormente il confronto tra i due giochi, al fine di precisare la struttura del gioco dell'interazione e far finalmente giocare il robot umanoide.

Possiamo innanzitutto notare che il modello della percezione fornito dal *predictive engagement* pone il gioco dell'interazione all'interno di una dimensione composta di abiti e aspettative *embodied*. Diversamente da quanto prospettato nel modello del *predictive coding*, il processo di costruzione di abiti e aspettativi che si manifestano in interazione si connota di una dipendenza costitutiva i) del corpo, ii) delle esperienze precedenti che hanno coinvolto il corpo attivamente e hanno contribuito alla costruzione di abiti sensorimotori come pattern enattivabili in concomitanza di circostanze simili e iii) della qualità concreta dell'interazione in atto, che indirizza il processo di costruzione di senso concedendo la realizzazione di alcuni abiti e non altri, la soddisfazione di alcune aspettative e non altre. In sostanza, se riportiamo il discorso al gioco dell'interazione dell'HRI, la H in questa sigla, lo *Human*, si presenta all'incontro con un portato esperienziale definito, composto dalle abitudini alle interazioni soggettive realizzate nel corso della sua storia e rinnovabili all'occorrenza della prova interazionale. Le aspettative, in questa struttura abituale, sono i risultati della ripetizione con variazione dell'abito interazionale sensorimotorio enattivato, il legame con il quale forma quella struttura malleabile ma robusta di inclinazioni ad agire in modo simile in circostanze simili e la capacità di riconoscerne l'efficacia pragmatica nel contesto di enattivazione. Nel confronto con il gioco dell'imitazione, possiamo registrare questo aspetto come elemento in comune con il gioco dell'interazione: sebbene non reso esplicito, anche nel caso dell'*imitation game* l'essere umano valuta le risposte della macchina sulla base delle esperienze precedenti maturate in altre conversazioni con esseri umani. Solo così lo sperimentatore può essere capace di accorgersi quando il suo interlocutore è una macchina e non un essere umano, ovvero solo nell'intrico di abiti interazionali e aspettative

dialogiche, che anche nel caso del linguaggio sono – almeno in parte – incarnate (Di Paolo et al. 2018; Cuffari et al. 2015).

Allora, ciò che è importante sotto il profilo delle predizioni *embodied* per il gioco dell'interazione concerne almeno due aspetti: da una parte, il *predictive engagement* si declina come la struttura soggiacente alle capacità dell'agente umano di imbastire dinamiche fluide che pescano dal passato degli abiti e gettano un arco nel futuro delle aspettative: il presente si pone come momento di raccordo tra queste due dimensioni, l'unica che mostra uno stato di concretezza tangibile come il nodo in cui le dimensioni temporali, virtuali o realizzate, si concatenano a formare il piano percettivo dell'agente; dall'altra, il *predictive engagement* conferisce al robot un'occasione, ovvero la possibilità di giocare al gioco dell'interazione mostrando la robustezza dell'aspetto sincretico: nella danza che stiamo osservando, il robot può sicuramente presentarsi sulla pista da ballo.

Tuttavia, il gioco non è ancora iniziato. Possiamo osservare che moltissime delle ipotesi che abbiamo analizzato per spiegare l'UVE si fermano a questo punto del gioco, credendo che il robot mostri già alla sua manifestazione le caratteristiche per non poter partecipare. Questo è testimoniato dal fatto che, nella quasi totalità degli esperimenti e delle ipotesi annoverati, il modo in cui viene testato l'UVE è attraverso la somministrazione di immagini virtuali e fotografie, ovvero rappresentazioni statiche che non concedono al robot nemmeno il beneficio del dubbio di poter esprimersi a livello interazionale. Eppure, nella nostra analisi dell'aspetto sincretico, il robot sembra poter giocare eccome, muovendosi sulle note di quell'armonia tra mimesi e sintesi che ne compone la manifestazione sincretica. Possiamo, senza riserve, dare il via alle danze.

Grazie all'approfondimento reso possibile dal PSM, siamo capaci ora di specificare la forma del gioco, ovvero quella della coordinazione e della individuazione del ritmo interazionale. Nel reticolo allestito da abiti e aspettative, le posizioni non sono tutte già occupate: è, infatti, nello spazio tra i gesti, in quel vuoto che emerge fra un movimento e il prossimo, che l'altro è chiamato a partecipare; è attraverso il riconoscimento del vuoto da riempire, della pausa fra le azioni che chiede l'intervento dell'altro, che è possibile giocare insieme. Per riempire le posizioni allestite dall'interazione, abbiamo visto che è necessario saper coordinare i propri abiti sensorimotori a quelli esibiti nel comportamento dell'altro, così soddisfacendo le aspettative che sia l'altro, sia l'interazione stessa pongono a condizione dell'efficacia pragmatica. L'esempio della danza, qui, è decisamente euristico: il modo attraverso cui è possibile realizzare forme di coordinazione tra due agenti sulle note di una musica necessita di una coordinazione sensorimotoria che tenga conto sia del proprio corpo, sia del corpo dell'altro, sia dell'ambiente in cui la danza viene realizzata.

Players/dancers have to engage with what is presently at hand [...and] respond to the different possibilities that appear in the moment. This means that sometimes you move along with the suggestion of another participant, while at other times you actively sculpt

and re-direct the situation. To respond creatively to the moment, the players/dancers need to attune to their own bodies, other bodies and surrounding. They need to be sensitive to the relational parameters and open to the creative potential that might pop up. (Hermans 2023: 5)

Da queste forme di complesso aggiustamento in itinere, dunque, si produce come risultato un elemento terzo, l'interazione, che si configura a sua volta come il terreno sdruciolevole dell'organizzazione del significato: se la coordinazione è mantenuta e gli agenti si agganciano al ritmo esibito e rintracciato, allora il risultato sarà armonico, fluido, familiare come quello che può assumere l'incontro con un'amica di vecchia data. Non si farà fatica a riempire gli spazi da lei lasciati vuoti per noi, perché la storia delle nostre interazioni passate parlerà al posto nostro: gli abiti sensorimotori si concateneranno senza quasi bisogno di aggiustamenti, sovrapponendosi e disgiungendosi all'occorrenza. Quel pattern sarà così familiare poiché ad abiti sensorimotori enattivati seguiranno costantemente aspettative corporee soddisfatte, posizioni opportunamente occupate e intrecci delle azioni in coordinazione, con il tipico spessore ritmico dell'amicizia e del mutuo riconoscimento.

D'altra parte, abbiamo anche visto che l'esito della coordinazione si può manifestare in momenti di rottura che, lungi dall'essere considerati esiti definitivi dell'interazione andata male, costituiscono il battito nel ritmo interazionale. Il *breakdown* è ciò che organizza la struttura del senso e consente di puntellare lo sviluppo dell'interazione, un elemento imprescindibile nel diagramma della significazione: la rottura del ritmo non è la sconfitta, ma parte del gioco, che anzi acquista la possibilità di evolversi nell'ottica compartecipata del reciproco riconoscimento. I tentativi di saldare le crepe dell'interazione sono opportunità che si attende vengano colte insieme, nell'intento di ricucire lo strappo generato dal *breakdown* e riassetare o rinnovare i panorami percettivi e cognitivi, sia quelli di ognuno degli agenti coinvolti, sia quello generato dall'interazione sociale.

Tramite questo aspetto, riusciamo a mettere in luce la vulnerabilità intrinseca dell'interazione, la quale continuamente si rompe e si ricompone. Ogni rottura non rappresenta solo un limite ma diventa uno spazio di riadattamento e di negoziazione del significato, in cui il senso stesso dell'interazione viene messo alla prova. D'altronde, abbiamo sottolineato già in §5.1.1 come nella quotidianità non si possa smettere di giocare al gioco dell'interazione, anche quando le rotture sembrano apportare un alto grado di entropia. È in questi momenti di discontinuità che emerge la vera natura dinamica del dialogo, poiché ogni ricostruzione incorpora le tracce dei passaggi precedenti, modellando l'esperienza futura. La vulnerabilità non è un punto debole che si mostra scoperto, ma un'opportunità per ridefinire il rapporto, per rinnovare le aspettative e riorganizzare le intenzioni in un ciclo dinamico e aperto. Quindi, grazie alla teoria dei micromondi di Varela, infine, possiamo esplicitare il fatto che il gioco dell'interazione è esattamente il processo di ricostruzione di micromondi a seguito delle rotture interazionali che si manifestano mentre stiamo giocando. Si tratta sempre di recuperare in qualche modo dallo sgambetto del *breakdown*, che momentaneamente ci obnubila ma che il *sense-*

*making*, in qualche modo, riorganizza e sistematizza. Ciò non significa che tutte le interazioni sociali possano essere recuperate, proprio perché il contributo dei singoli è necessario al processo di stabilizzazione del senso, pena il cambiamento della dinamica interazionale e la cessazione della socialità intrinseca al PSM.

#### 4.4.10.1 Giochi a confronto

Ora, prima di passare ad analizzare il modo in cui questo approccio possa rendere conto dell'UVE tramite il gioco dell'interazione che abbiamo descritto, dobbiamo procedere al confronto con il gioco dell'imitazione. Grazie alla rilettura che abbiamo proposto, infatti, possiamo sviluppare alcune riflessioni, che ci conducono a ribaltare il rapporto tra i due giochi: non è più il gioco dell'interazione a costituire una variante del gioco dell'imitazione, ma esattamente il contrario.

- In primo luogo, possiamo tracciare un parallelo tra le strutture predittive del modello del *predictive coding* e il gioco dell'imitazione per come è stato pensato da Turing. Come abbiamo notato in §5.1, l'errore nell'*imitation game* sancisce il momento di interruzione della prestazione: il riconoscimento della macchina come incapace di giocare *allo stesso gioco* della controparte umana mette la macchina nella posizione di definitiva sconfitta. Se la macchina non è in grado di imitare esattamente le stesse modalità di risposta che sarebbero proprie di un essere umano, l'errore registrato si manifesta non solo come un'incrinatura nell'interazione, ma come la sua chiusura risolutiva. Il processo, quindi, prevede l'intervento diretto sulla macchina, che deve essere modificata al suo interno affinché possa riuscire a mostrare, nei giochi successivi, più somiglianza all'umano, più capacità imitative, più abilità derivanti dalla modellizzazione computazionale del pensiero. In questo senso, il parallelo con il *predictive coding* è presto fatto: anche nel caso della registrazione dell'errore nel processo predittivo computazionale c'è un momento di arresto, di interruzione di fronte all'impossibilità di minimizzare l'errore di predizione, che sancisce, in un certo senso, il fallimento del processo percettivo. Anche qui, il meccanismo si ricalibra guardando all'interno della struttura computazionale al fine di allineare l'esito della percezione con i suoi presupposti e far sì che nel prossimo processo predittivo il modello sia più performante, più capace di anticipare il *mismatch*, più efficace nella minimizzazione della sorpresa affinché non si ritrovi spiazzato anche la prossima volta. Tuttavia, questo tipo di modellizzazione non considera il ruolo attivo che i processi sensorimotori e le pratiche incarnate assumono sia nel processo di rilevazione dell'errore, sia nello sviluppo dell'interazione, la quale continua senza necessità di auto-osservazioni dirette all'interno della struttura della mente. Entrambi questi modelli, infatti, trattano l'errore come un evento isolato e circoscrivibile, astratto dai processi di aggiustamento

incarnati e non imbrigliato nelle dinamiche dello sviluppo dell'interazione. Come il *predictive coding* sta al test di Turing, così il *predictive engagement* sta al gioco dell'interazione: nel gioco che proponiamo, mentre il *predictive coding* riduce l'errore a un'informazione da correggere puntualmente per mantenere una performance predefinita, il *predictive engagement* si orienta verso l'esperienza incarnata e situata, dove incertezze e pause diventano occasioni per il riadattamento del micromondo, senza che sia necessario riportare il sistema a uno stato di ottimizzazione funzionale, ma cercando delle situazioni sub-ottimali (Varela et al. 1991) in cui è possibile continuare a far scorrere il flusso del senso.

- In secondo luogo, possiamo nuovamente riconsiderare il ruolo che il *breakdown* assume nei giochi, alla luce delle riflessioni che hanno sottolineato l'importanza della processualità dell'interazione. Nel gioco dell'interazione che proponiamo, infatti, le discontinuità non sono difetti da eliminare, ma elementi vivi che sollecitano l'aggiustamento e la costruzione di nuovi significati, rendendo l'interazione stessa un processo aperto e co-creativo. In questo senso, la teoria dei micromondi di Varela ci consente di inserire il *breakdown* in uno sviluppo progressivo dell'interazione – anche sociale – il quale non prevede che l'incontro si disgreghi sotto il peso della rottura interazionale. Se riprendiamo l'esempio della stretta di mano mancata, questo aspetto sembrerà ovvio: anche se molto delusi e, magari a disagio, di fronte al diniego del gesto sociale che abbiamo offerto non possiamo congelarci, non interagire più, o scappare, appunto, a gambe levate. Dobbiamo continuare ad abitare quell'ambiente, anche se per poco, e in quel frangente portare avanti qualche forma di interazione. Nei termini del gioco dell'interazione, questo significa rimanere sul terreno di gioco aperto dall'incontro con l'altro, anche se oramai staremo ballando da soli. Nei termini della teoria dei micromondi, questo implica il riallestimento delle salienze percettive, che veicolano nuovi significati pragmatici: potremmo essere più attenti al segnale che un nostro collega ci fa dall'altra parte della stanza e prendere al volo l'occasione di fuggire da quell'incontro andato male; potremmo voler cercare qualcun altro con cui parlare, per toglieroci da quel momento imbarazzante. In ogni caso, il nostro piano percettivo sarà riconfigurato e i nuovi micromondi forniranno nuove possibilità di significazione, nuovi ganci a cui concatenarsi, nuovi comportamenti di cui individuare il ritmo. È questa la struttura che si delinea per il nostro gioco dell'interazione, che sembra molto differente da quella del test di Turing. Nel gioco dell'imitazione, abbiamo notato che l'errore non deve essere recuperato insieme, e che anzi il gioco sia giocato praticamente in solitaria: il parallelo più utile è con quei giochi come il braccio di ferro, in cui entrambi gli agenti devono realizzare lo stesso compito, in competizione l'uno contro l'altro, e non in collaborazione. Alla fine, uno dei due risulta sconfitto: la macchina perde la possibilità di

essere considerata intelligente se l'essere umano la scopre; l'essere umano perde la sua unicità nel mondo delle creature intelligenti se la macchina non viene scoperta. Il gioco finisce comunque, e termina con vincitori e perdenti che fanno i conti con le proprie debolezze. Invece, il gioco dell'interazione è più simile alla danza: un gioco in cui si entra in pista, la musica si accende ed entrambi i giocatori hanno un obiettivo in comune, che è performare. Il compito prevede, come abbiamo visto, processi di mutui intendimenti, e allo stesso tempo non si può smettere di ballare all'improvviso, solo perché l'altro ha sbagliato un passo della coreografia o perché la musica non è quella giusta. Alla fine, non ci sono sconfitti e vincitori, ma solamente il risultato dello spettacolo, che sarà il pubblico a giudicare.

Da questo paragone, ciò che possiamo evincere è un ribaltamento delle prospettive, alimentato dal fatto che l'imitazione al centro del test di Turing non sia altro che una manifestazione di processi di *mimesi* più ampi (Dumouchel & Damiano 2011), che includono anche l'imitazione come una delle possibili declinazioni della coordinazione, la quale è anche sincronizzazione, *mirroring* e anticipazione (De Jaegher & Di Paolo 2007: 490), oltre che alle altre modalità attraverso cui è possibile concatenarsi nel ritmo interazionale. La struttura del gioco dell'interazione, quindi, ci permette di includere molti più elementi, e così giocare in un campo più ampio, con maggiori possibilità di confronto e con uno sguardo privilegiato ai passaggi tra rotture e riconfigurazioni dei micromondi che nel test di Turing non vengono considerati.

Ora che abbiamo esplicitato i meccanismi del gioco dell'interazione, è il momento di far giocare il robot. Abbiamo individuato lo scopo del gioco nella comprensione della misura in cui il robot riesce a interagire con l'essere umano e, a seguito della nostra disamina dell'UVE, possiamo includere anche l'obiettivo di comprendere a che livello dell'interazione si presente questo effetto.

#### 4.4.10.2 *La danza dell'interazione: analisi dell'UVE agli occhi di chi guarda*

Attraverso gli strumenti concettuali che abbiamo elaborato, allora, possiamo scomporre il gioco in tre fasi: i) la fase delle aspettative, ii) la fase dell'*engagement* e iii) la fase del micromondo alternativo.

i) *Fase delle aspettative*. Di questa fase abbiamo parlato a lungo, e abbiamo specificato nei dettagli la struttura delle aspettative richiamate dagli abiti sensorimotori convocati nel micromondo delle interazioni sociali. In questa fase, possiamo tracciare il profilo di come l'essere umano e il robot si presentano all'incontro dell'interazione, ognuno con i suoi portati – esperienziali e computazionali –, ognuno con le sue capacità. Per quanto riguarda l'essere umano, il portato esperienziale consiste nelle specifiche e individuali capacità di integrazione dei propri abiti all'interno del tessuto interazionale, i quali comportano l'enattivazione delle aspettative pragmatiche della coordinazione

sensorimotoria tipiche del micromondo della socialità, di cui abbiamo parlato in §5.3.1-5.3.2. Prendendo in considerazione il robot, d'altra parte, il suo portato si manifesta: a) nella sua capacità mimetica esibita tramite la somiglianza all'umano, similarità che non si scorge solamente come caratteristica statica della sua materialità ma si esprime anche in una forma di vita artificiale, mimando non solo le fattezze estetiche ma anche il rapporto agentivo istituito tra corpo e ambiente; b) nelle sue capacità sintetiche di riproduzione del comportamento umano, nelle molte forme che la tecnica robotica è riuscita a implementare: il movimento nell'afferrare un oggetto, la direzionalità dello sguardo, la gestualità conviviale impressa nel moto delle mani, della testa, delle labbra, ma anche le capacità verbali, quando presenti, che possono instradare un dialogo. Questo è il momento in cui le capacità si presentano insieme, simultaneamente, a comporre quell'armonia degli effetti che abbiamo individuato come il suo aspetto sincretico: causa ed effetto di un *anello interazionale* in cui gli esiti non possono essere compresi se non passando per la complessità delle relazioni potenziali e attuali incluse nell'incontro con gli altri. Come abbiamo visto, non è a questa fase che si può ricondurre l'UVE come rottura interazionale. Gli agenti sono pronti a giocare.

ii) *Fase dell'engagement*. In questa fase, la relazione si fa viva e il robot viene messo alla prova del gioco dell'interazione. Nel percorso che abbiamo compiuto, è stato più volte sottolineato che il processo di costruzione di senso si sviluppa nel controbilanciamento di fasi di coordinazione, in cui il micromondo appare definito e le *affordance* richiamano l'azione enattivando significati pragmatici, e fasi di *breakdown*, rotture interazionali che possono essere recuperate da entrambi gli agenti in interazione oppure in solitaria, rompendo quindi quella socialità cifra distintiva del PSM. Abbiamo escluso, d'altra parte, che l'UVE possa presentarsi come semplice *breakdown* dell'interazione, dal momento che la realizzazione di molti altri tipi di rotture non comporta una sensazione di disagio come quella ritenuta caratteristica dell'effetto perturbante. Cosa accade, quindi, in questa fase? La scena che vogliamo descrivere fa riferimento esattamente al circuito istituito da abiti sensorimotori, aspettative relazionali e rotture della coordinazione. Per comprendere meglio il meccanismo che stiamo descrivendo, è sufficiente pensare a cosa accade in una qualsiasi conversazione<sup>191</sup>: quando parliamo con qualcuno, il modo attraverso cui si concatenano gli enunciati dà vita a un flusso conversazionale che può prendere anche pieghe inaspettate (si pensi all'esempio della discussione involontaria con il nostro compagno, di cui abbiamo parlato poco sopra) o manifestarsi in incomprensioni e richieste di chiarimenti. Tuttavia, in una conversazione, non è solamente questo il livello su cui si istituisce il tentativo di coordinazione: tutti i processi che abbiamo elencato facenti

---

<sup>191</sup> La conversazione costituisce un esempio emblematico di un possibile scambio interazionale, presentando oltretutto una struttura schematica il cui diagramma è facilmente estraibile. In questo modo, crediamo si possa al meglio mostrare quale sia la dinamica interazionale in cui aggiustamenti, coordinazioni e rotture si sviluppano in sequenze autonome di costruzione di senso condiviso.



parte delle operazioni di mutuo riconoscimento, quali la sincronizzazione, il *mirroring* e la mimesi, sono allo stesso tempo costitutivi della capacità di coordinazione degli agenti, in un processo che corre almeno sul doppio binario dello scambio verbale e dell'*attunement* corporeo<sup>192</sup>. Quindi, oltre a rotture che potremmo definire esplicitamente rintracciabili – una incomprensione linguistica, interruzioni nella comunicazione o problemi nel passaggio del suono –, ci sono altri tipi di rotture, che per noi risultano ancora più interessanti, dato che avvengono al livello della coordinazione motoria: *breakdown* dati dalla mancanza di *attunement* solitamente realizzati nelle circostanze dell'interazione sociale, come ad esempio il *mirroring* delle espressioni facciali. Diversi studi, ad esempio, testimoniano dell'importanza del *mirroring* come segnale corporeo di reciproca comprensione e riconoscimento intersoggettivo anche al primo incontro (Navarretta 2014, 2016), supportato dal sistema neurocognitivo dei neuroni specchio (Rizzolatti 2005; Rizzolatti & Fabbri-Destro 2008). Come illustrato da Iacoboni (2009: 120):

By firing as if were actually making those facial expressions we are simply observing, these neurons provide the mechanism for simulated facial feedback. This simulation process is not an effortful, deliberate pretense of being in somebody else's shoes. It is an *effortless*, automatic, and unconscious inner mirroring.

Dunque, in contesti di incontro intersoggettivi, quello che accade è che la coordinazione avviene su più livelli, di cui quello dell'*attunement* corporeo e sensorimotorio è la base su cui le altre dimensioni si costituiscono. In sostanza, nell'interazione sociale, la nostra comprensione di ciò che sta accadendo, che include il richiamo degli abiti sensorimotori adeguati e la formazione di aspettative correlate, passa per una percezione delle espressioni dell'altro come *affordance* che consentono lo sviluppo del ritmo condiviso. Situazioni in cui ciò non avviene, ovvero in cui non rintracciamo questo *attunement* corporeo-percettivo reciproco, sono causa di disagio e, a volte, di forte stress. Una celebre serie di esperimenti (Tronick et al. 1989; Murray & Trevarthen 1985) mostra come per bambini in tenera età la mancata coordinazione motoria che si realizza nel *mirroring* e nelle altre forme di allineamento interazionale nelle interazioni con la madre conduca a esperire una forte sensazione di disagio. Ad esempio, nel celebre esperimento di Murray e Trevarthen (1985), quattro bambini di età compresa fra le 6 e le 12 settimane di vita sono stati osservati interagire con la loro madre attraverso uno schermo. Lo scenario prevedeva due versioni: la condizione *live*, in cui l'interazione tra madre e bambino era in tempo reale, con scambi reciproci che si verificavano senza ritardo, e una versione *tape*, in cui ai bambini veniva presentata la registrazione del comportamento della madre. Murray e Trevarthen riportano nei risultati che i bambini si mostravano altamente collaborativi nella interazione

---

<sup>192</sup> È risaputo, ad esempio, che la comunicazione non verbale (microespressioni facciali, assetto corporeo, ecc.) nei contesti conversazionali ricopre un ruolo fondamentale per la costruzione del processo interazionale, insieme ad altri fattori che non pertengono alla struttura linguistica, come ad esempio il tono di voce, le sfumature emotive, il ritmo dell'interazione (cfr. Burgoon & Bacue 2003).

con la madre nella condizione *live*, mentre nella condizione *tape* i bambini esprimevano disagio e malessere, molto spesso piangendo, o perdevano interesse nell'interazione stessa. Anche nelle elaborazioni successive (Rochat et al. 1998; Adamson & Frick 2003), questo risultato è stato interpretato in relazione alla capacità del bambino di comprendere, già a partire dai 2 mesi di vita, le microespressioni prodotte dalla madre in interazione come un segno dell'*attunement* con lei. Invece, nei casi in cui si trovavano improvvisamente ad avere a che fare con la registrazione del comportamento della madre, o con fermi immagine del suo volto (*still-face condition*), i bambini mostrano reazioni cariche di stress (come *gaze aversion*, *distancing behaviors*, e *autonomic stress indicators*, cfr. Adamson & Frick (2003), o anche «crying, yelling, hitting the mother, or throwing toys» Weinberg et al. 2008: 5)). Questo tipo di reazione è stato messo in relazione con altri tipi di stress registrabili nella relazione con la madre, sottolineando che «infants appear to find it more disturbing than several other perturbations of social interactions, including brief physical separations» (Adamson & Frick 2003: 460). Inoltre, altri studi sottolineano che i bambini, nei momenti immediatamente successivi alla interruzione dell'interazione data dalla presentazione della registrazione o del fermo immagine, mostrano espressioni come sorrisi, sbadigli e movimenti del corpo accentuati, ipotizzati essere gesti di attivazione dell'*attunement* interazionale, basati sulla – seppur breve – esperienza dell'agente (come ad esempio *pick-me-up gestures*, mostrando giocattoli alla madre, cercando attirare la sua attenzione con vocalizzi o «repeating verbal utterances, which became increasingly loud and more insistent when the mother did not respond» (Weinberg et al. 2008: 17)).

Non si farà fatica a riconoscere in questi casi una struttura di controbilanciamenti tra rotture interazionali e coordinazioni nell'*engagement* fra gli agenti: una volta che i bambini capiscono che c'è stato un *breakdown* dell'interazione, tentano di recuperare attirando l'attenzione della madre, mostrando la loro condizione emotiva in modo esplicito, rendendo di fatto palesi le loro aspettative e sviluppando comportamenti che rivelano la loro frustrazione e il loro malessere. Questo tipo di processi ci mostra non solo che l'*attunement* emotivo fatto di *engagement* interazionali basati sulla coordinazione e sulla rilevazione del ritmo è una componente fondamentale del nostro modo di comprendere gli altri (De Jaegher & Di Paolo 2007; Di Paolo et al. 2018; Gallagher 2005, 2021), ma anche che la loro rottura, generata non da un solo disallineamento relazione, ma da numerosi tentativi di recupero falliti, può sfociare in sensazioni di forte disagio, che mettono l'agente in una condizione interazionale non confortevole. Questa serie di esperimenti e la loro collocazione all'interno del gioco dell'interazione ci consente di specificare cosa succede nel caso dell'UVE.

La nostra proposta, infatti, consiste nel concettualizzare l'UVE come un fenomeno che avviene proprio nella fase dell'*engagement*, non come la rottura dell'aspetto interazionale con l'essere umano,

ma generato dall'incapacità del robot di coordinarsi *nel corso dell'interazione* in termini di abili sensorimotori e, quindi, mancare la presa con le aspettative incarnate che vengono richiamate nel micromondo dell'interazione sociale. Questo avviene in un processo multidimensionale, che incrementa di complessità in dipendenza delle capacità esercitabili dal robot nel contesto interazionale, il cui minimo denominatore, d'altra parte, risiede nell'*attunement* sensorimotorio realizzato nelle coordinazioni delle microespressioni del corpo e nei fenomeni di *mirroring*, di sincronizzazione e di mimesi così fondamentali nella coordinazione intersoggettiva. Nonostante le eccezionali caratteristiche estetiche che un robot con alta somiglianza all'umano possa mostrare, queste – almeno ad oggi – non sono correlate da capacità espressive continue, dinamiche e plastiche che contraddistinguono le modalità espressive di coordinazione tra umani: ciò conduce alla comparsa di numerosi *breakdown* interazionali, tanti più quanto il tempo dell'interazione si prolunga, fino a portare all'emersione dell'effetto di straniamento tipico dell'UVE.

In particolare, ciò che accade è che, nell'intessersi della trama di sincronismi attesi, l'essere umano allestisce percorsi abituali della coordinazione sensorimotoria. Il robot, d'altra parte, non avendo la capacità fornire un'espressività tale da mostrare segni di coordinazione motoria adeguata e di comprendere genuinamente, a livello corporeo, gli abili incarnati del suo partner interazionale, si concatena alle azioni 1) solo se riesce a riconoscerle e 2) attraverso le modalità interazionali fisse, e non dinamiche, registrate nel suo sistema di competenze. Dal momento in cui il robot non può produrre tali forme di coordinazione, non si concatena come l'interazione vorrebbe alle dinamiche di organizzazione enattivate, realizzando così un *microbreakdown* dell'interazione. Se preso isolatamente, tuttavia, questo non dovrebbe essere grave: come in molte altre occasioni di rottura interazionale, la costruzione del senso condivisa ora si riassetta verso l'obiettivo di tenere in piedi l'interazione. Così, il micromondo cambia, e con lui gli abili incarnati richiamati al suo interno e le aspettative che ne derivano. Tuttavia, il robot, se non è riuscito a rendere conto della coordinazione di un'azione instradata nel percorso comune del senso sociale, supportato dalle norme condivise e dalla cultura di riferimento, difficilmente sarà capace di recuperare una rottura interazionale e di recuperare insieme la socialità del micromondo nella HRI. Questo è il gioco dell'interazione, d'altra parte, quello in cui si attraversano micromondi cavalcando microfratture interazionali, capacità che è data all'essere umano dalle evoluzioni delle forme di *sense-making* tipiche dell'organismo biologico complesso e dalla sua storia situata ed *embodied* all'interno di un mondo che è già sociale, che è già popolato da altri agenti che ne instradano le abilità. Il robot, quindi, non può che rimanere a guardare, macchina cibernetico-computazionale impegnata nel processo di computare le proprie operazioni e *sperare* di riorganizzare l'interazione.

L'UVE, quindi, sopravviene come effetto dell'incapacità di aggiustamento e coordinazione relazionale, mostrandosi come la somma non lineare dei *microbreakdown* che normalmente, in altre interazioni sociali, verrebbero in qualche modo recuperate insieme all'interlocutore. Affinché questa circostanza si realizzi, non c'è bisogno nemmeno che i due agenti parlino l'uno con l'altro o che debbano trovarsi a interagire all'interno dello stesso ambiente. È sufficiente che l'essere umano si trovi a osservare il robot interagire con altri agenti, o muoversi all'interno di un ambiente: questo basterà a innescare quel meccanismo di abiti e aspettative che renderà il robot, prima, possibile partecipante al gioco dell'interazione e, poi, elemento di disagio che disattende le aspettative dell'interazione. Questo perché il sistema di comprensione sensorimotoria enattivato nell'essere umano si sviluppa sempre su una dimensione prima di tutto *embodied*, in quel processo che abbiamo definito di *intentionality detection* (§4.4) che individua il pattern significativo dell'azione nella considerazione del rapporto che l'agente intrattiene con il suo ambiente.

Crediamo che l'UVE possa essere spiegata esattamente in questo senso, ovvero facendo appello alle dinamiche sensorimotorie normalmente instradate dall'interazione sociale, le aspettative derivanti dalle quali vengono più volte disattese nel corso dell'interazione con il robot e danno vita, progressivamente, al senso di inquietudine e disagio tipico dell'effetto *uncanny*. L'idea alla base di questa spiegazione trova risonanza in un articolo che indaga le reazioni di fronte all'androide Geminoid (Nishio et al. 2007; Ishiguro & Dalla Libera (Eds.) 2018; Ishiguro 2020) e mette in relazione le regolari attività di coordinazione nell'interazione intersoggettiva con l'esperienza interazionale con il robot:

*Geminoid* does not move enough. It is not just a question of basic physiological motion, the fact that a human's chest moves as he or she breaths, that our eyes blink all the time, our nostrils dilate, pupils and eyes change shape as we look at something far or close and depending on the quantity of light in the environment. Rather, it is that humans, unlike *Geminoid*, cannot stand still. They move all the time, scratch themselves, move their shoulder, turn their head, cross this leg then that, shift their weight from left to right or front to back. These, and an innumerable quantity of others movement are mostly unconscious, we are unaware of them, and we are mostly unaware of similar movements of the part of the interlocutors. Yet, even if we are unaware of them, they form a fundamental part of the experience of being in the presence of another human. *Geminoid* does not move except when it engages in an explicit motion, among us explicit motions arise out of a continuum of movements that only ends with death. These movements further play a fundamental role inasmuch as they are not simply random. When two individuals are in presence of each other, their body motions tend to become *coordinated*. Among us, there is a fundamental mimesis that takes place at the level of these simple spontaneous and mostly unconscious motions. If the person you are talking to touches her face the probability that you will touch your face in the next minute goes up dramatically. This basic bodily mimesis, this physical conversation that is permanently ongoing between interlocutors is part of the way in which we collectively determine each other's intentions toward each other. This *Geminoid* does not do. It is not part of this exchange. (Dumouchel & Damiano 2011: 30)

Purtroppo, ad oggi non è possibile fare appello a esperimenti o casi studio per validare in modo concreto questa ipotesi, dal momento che nessun esperimento per validare l'UVE è stato mai svolto in presenza (Paetzel & Castellano 2019), ne è mai stata considerata la componente corporea e situata in interazione con il robot come costituente fondamentale nell'emersione del fenomeno. Tuttavia, possiamo richiamare una serie di esperimenti relativi (a) alla differenza di manifestazione del fenomeno in gruppi di persone con diversi *background* culturali, e (b) agli studi sulla persistenza dell'UVE in interazione con il robot.

- a) Con l'accrescimento di interesse nei confronti dell'UVE, alcuni studi (Bartneck et al. 2007; MacDorman et al. 2009; Kaplan 2004; Castelo & Sarvary 2022; Rosenthal-Von Der Pütten et al. 2014; Destephe et al. 2015) hanno indagato le possibili differenze nella manifestazione dell'UVE, principalmente differenziando i gruppi in due macroaree: Oriente e Occidente globali. In generale, si può notare che i risultati degli studi sulle differenze culturali nell'UVE siano piuttosto controversi, dal momento che alcuni rintracciano differenze sostanziali nell'emersione del fenomeno e quindi nella possibilità di interagire con robot umanoidi (Bartneck et al. 2007; MacDorman et al. 2009; Kaplan 2004; Castelo & Sarvary 2022; Destephe et al. 2015), mentre altri non rilevano differenze significative tra popolazioni appartenenti all'est del mondo o all'occidente (Rosenthal-Von Der Pütten et al. 2014). Approfondendo la comparazione fra gli studi, tuttavia, più volte è sottolineato che individui con *background* culturale di tipo occidentale (in questo caso, statunitensi) sembrano essere più a loro agio con robot che mostrano bassi livelli di somiglianza all'umano, mentre accade il contrario per individui provenienti dal Giappone (cfr. Castelo & Sarvary 2022: 1871). Lo stesso, espresso in diversa maniera, è affermato da Destephe e colleghi (2015), i quali hanno riscontrato una differenza sostanziale nella valutazione di parametri come *eeriness* e *attractiveness* nei questionari compilati da giapponesi e francesi. Nell'esposizione delle motivazioni chiamate a giustificare questa differenza, gli autori affermano:

Japanese people and French people have different views on what is natural and artificial [...]. French people see natural things and artificial things as opposed: they see the world as hierarchical, boundaries limiting things and categories defining them. This mindset might be influenced by a Cartesian French education [...]. Inheriting a tradition of Buddhist (everything is considered to be a manifestation of the same greater concept) and Shintoism (spiritual essence can be manifested in any form from rock to rivers through animals and even humans), Japanese people would see natural things and artificial things connected and being parts of a bigger picture. These distinctions might influence the Eeriness perception: Japanese participants would be less sensitive to discrepancies in the robot, thus rating lower Eeriness than French participants. Japanese people will prefer Natural Intensity emotion representation and French people will prefer Exaggerated Intensity emotion representation. (Destephe et al. 2015: 2)

In sostanza, le differenze nell'esperienza della *eeriness* sono ricollegate a una serie di conoscenze precedenti, rintracciabili nelle diverse modalità di vedere il mondo: da una parte, il meccanicismo riconducibile all'impronta cartesiana per l'Occidente (cfr. §1.1.2), dall'altra la visione olistica tipica delle culture e delle religioni orientali per l'Oriente. Avevamo visto che questa differenza fosse importante anche per la percezione del robot come agente intenzionale nel capitolo §4 e come le differenti *propositional attitude* venissero richiamate e intrecciate a forme di percezioni corporee di intenzionalità. Quello che vogliamo sottolineare ora è che, oltre a un portato culturale differente sulla visione del mondo, anche il modo in cui le diverse culture affrontano le interazioni sociali potrebbe essere un fattore da considerare nell'esperienza dell'*uncanniness*. Infatti, una intera corrente di studi antropologici, sociali e cognitivi (Huang & Deng 2008; Uono & Hietanen 2015; Qingxue 2003; Aketchi et al. 2013; Senju et al. 2013) mostra che i pattern sociali variano tra diverse culture e che questo potrebbe influire notevolmente sui modi in cui entriamo in interazione con l'altro. In altre parole, detta nei termini del modello del PSM, a differenze culturali potrebbero corrispondere diverse possibilità di concatenazione nell'interazione, diverse modalità di rintracciare i ritmi interazionali, e quindi differenti modalità per il robot di disattendere le aspettative corporee aperte dall'interazione. Inoltre, abbiamo visto che anche le pratiche percettive subiscono dei condizionamenti in base ai portati culturali (cfr. Chua et al. 2005; Masuda & Nisbett 2001; Winerman 2006) e questo potrebbe influire notevolmente sulle modalità di emersione degli *affordance* sociali. È stato rilevato, infatti, che pratiche percettive come quelle di contatto visivo sono molto diverse fra orientali (giapponesi) e occidentali (finlandesi), portando anche a diverse interpretazioni delle espressioni facciali e all'interpretazione delle emozioni dell'altro: i finlandesi tendono a sostenere di più lo sguardo dell'altro e «to consider a slightly averted gaze as directed at them when viewing Japanese faces than when viewing Finnish faces» (Uono & Hietanen 2015: 13). Le motivazioni sono ricondotte dagli studiosi al fatto che in occidente è molto più frequente l'esperienza dello sguardo diretto dell'altro, mentre nella cultura orientale si preferisce non incrociare spesso lo sguardo dell'altro, soprattutto in contesti formali o non familiari. In sostanza, le nostre ipotesi ci spingono a considerare tra i fattori di influenza dell'UVE anche le differenti modalità interazionali intersoggettive: mentre in occidente l'interazione sembra essere molto più diretta, includere anche momenti di contatto e basarsi su una coordinazione spontanea nel processo di aggiustamento reciproco, nel caso di culture le cui interazioni sociali sono molto più formali, basate su regole rigide e su una distanza corporea che coinvolge anche lo sguardo questo comporterebbe delle diverse modalità di aggiustamento e delle diverse aspettative, che potrebbero rivelarsi molto più in

linea con le rigide modalità di interazione di cui un robot risulta ad oggi capace. Infatti, diversi studi mostrano come gli effetti di straniamento e inquietudine di fronte ad androidi o robot umanoidi non sia così alta in Giappone, paese con uno dei più alti gradi di presenza di robot altamente somiglianti all'umano anche in contesti quotidiani, ma anche con regole sociali molto diverse dalle nostre che si basano su una etichetta relazionale precisa e che coinvolgono il corpo in maniera molto minore rispetto alla cultura occidentale (cfr. Nisbett 2003).

- b) Un gruppo di studi mostra come l'UVE non svanisca dopo il primo incontro con il robot, ma persista anche nelle interazioni successive, anche se sovente si affievolisce (Strait et al. 2015, 2017; Zlotowski et al. 2015; Alvarez Perez et al. 2020). In particolare, uno dei primi studi elaborato su questo tema è quello di Zlotowski e colleghi (2015), i quali ipotizzano che «eerie robots could produce affective habituation and the initial strong negative emotional response will weaken with increased exposure» (2015: 3). Questo esperimento è di particolare interesse, poiché il setting sperimentale prevede che l'essere umano e il robot, anche in questo caso l'androide Geminoid, siano nello stesso luogo e possano interagire, oltretutto in più sessioni e in un arco di tempo prolungato. I risultati registrati dai ricercatori mostrano che «a robot that is perceived as uncanny is not able to affect its likeability by a positive or negative interaction» e che «after the last interaction [the robot was] perceived as less eerie than after interacting with [it] for the first time» (Zlotowski et al. 2015: 9). Questo implica che l'esposizione prolungata in interazione con un robot umanoide sembra suscitare dapprima l'UVE maggiore, con una risposta emotiva legata all'inquietudine al suo massimo, per poi affievolirsi nel corso del tempo. Allo stesso tempo, comunque la sensazione non scompare, e lo stesso è confermato da esperimenti successivi (Paetzel & Castellano 2019; Paetzel et al. 2020). Questo fattore ci conduce a riconsiderare l'ipotesi della VEH: se la spiegazione fornita tramite l'adesione alle teorie del *predictive coding* fosse giusta, il *mismatch* percettivo condurrebbe alla ricalibrazione del meccanismo di minimizzazione di predizione dell'errore, che nelle occorrenze successive non dovrebbe più registrare alti livelli di sorpresa alla somministrazione dello stimolo visivo del robot. Questi risultati, invece, sembrano confermare una interpretazione che considera i processi di coordinazione sensorimotoria e affettiva in interazione, che possono affievolirsi rientrando nella costruzione degli abiti e delle aspettative incarnate, ma si costruiscono nel tempo e distribuiti sul periodo delle interazioni in atto. Per questo, il robot si mostra comunque come *uncanny*, ma gli abiti sensorimotori enattivati e le aspettative a questi legate si nutrono delle esperienze incarnate precedenti e utilizzano l'interazione in atto come ulteriore prova dell'interazione, come ulteriore tentativo di enattivare il micromondo interazionale sulla scorta, d'altra parte, degli accoppiamenti di coordinazioni precedenti.

In sintesi, quello che reputiamo utile per riuscire a capire il motivo dell'UVE come effetto dell'interazione tra umani e robot consiste nella considerazione delle dinamiche incarnate dell'interazione sociale, che costituiscono il modo in cui l'essere umano si relaziona al robot e il terreno del gioco dell'interazione stesso, che può essere sfruttato solamente se entrambi gli agenti comprendono i ritmi e le occasioni di coordinazione che l'incontro stesso allestisce.

A seguito di questa analisi, possiamo quindi provare a rispondere alla domanda formulata nel gioco dell'interazione: *Can robots interact?* Se consideriamo i motivi per cui l'UVE si manifesta e che abbiamo elencato in questa sezione, il robot sembra non essere un buon compagno di giochi. Un ballerino provetto, per poter danzare con noi che lo facciamo da tutta la vita, deve saper intercettare i segnali sensorimotori che convogliamo nell'interazione, e il robot non sembra ancora capace di coglierne il significato profondo e agganciarsi ai nostri passi di danza, né di mostrarci esso stesso dei ganci interazionali che, come sottolineato da Dumouchel e Damiano (2011: 29), «form a fundamental part of the experience of being in the presence of another human». Tuttavia, non dobbiamo tralasciare un aspetto: il gioco dell'interazione non si ferma qui. Infatti, il momento di *breakdown* dell'interazione per come lo abbiamo concepito non decreta la fine dei giochi, ma costituisce un momento di raccordo: già nell'esperienza di rottura, si comincia a delineare il passaggio tra micromondi – fra quello che un attimo fa componeva l'orizzonte dei significati percettivi enattivabili e quello che potrebbe costituire il nuovo panorama di azioni da intraprendere. Nell'esempio fornito da Varela, la realizzazione di avere lasciato il portafogli sul bancone del bar è immediatamente seguita dall'organizzazione del micromondo che ne rende possibile il recupero: cosa succede, quindi, quando l'UVE costringe a riorganizzare il piano percettivo?

iii) *Fase del micromondo alternativo.* Questa fase si apre come prospettiva di risoluzione del *breakdown* interazionale aperto dall'UVE, come somma delle molteplici disattese alle aspettative incarnate che, in qualche modo, devono essere convertite in un nuovo assetto percettivo e interazionale, ovvero nel micromondo successivo. Per descrivere cosa accade in questa fase, dobbiamo innanzitutto analizzare cosa accade al robot nel nostro gioco. Se nella fase delle aspettative il robot era visto come potenziale interattante e interpretato a livello corporeo come un agente al pari dell'essere umano, la prova del gioco dell'interazione ha cambiato la sua situazione. Crediamo che il miglior modo per illustrare il cambiamento della condizione del robot sia quello di ricorrere, ancora una volta, alle categorie che abbiamo elaborato nel capitolo §3, attraverso cui abbiamo descritto il robot sotto l'aspetto di artefatto mimetico, di artefatto sintetico e di artefatto sincretico. Riprendendo questi strumenti, riusciamo a descrivere gli effetti dell'UVE sul robot come la rottura dell'aspetto sincretico. Ciò che risulta a seguito della realizzazione dell'UVE, infatti, è che il robot non viene percepito più come capace di concatenarsi agli abiti sensorimotori e di rispettare le aspettative del



suo interlocutore umano: come testimoniato dai questionari compilati a seguito dell'esperienza, il grado di realismo nei confronti di figure robotiche decresce rispetto ai momenti precedenti all'irruzione dell'effetto perturbante (MacDorman & Chattopadhyay 2016; Seyama & Nagayama 2007), per i robot androidi c'è un crollo della *animacy* percepita (Wang et al. 2020) e diventa più difficile a seguito dell'UVE riconoscere le emozioni espresse dai robot (Tschöpe et al. 2017). Detto nei termini del PSM, «interactions are social as long as the autonomy of the agents is not dissolved. If one agent becomes the sole regulator of the coupling, as in the use of a tool, this is no longer social interaction» (De Jaegher et al. 2010: 443): l'autonomia e l'intenzionalità sincretica che caratterizzavano il robot come artefatto si è frantumata sotto al peso delle incongruenze sensorimotorie e delle aspettative disattese, e mostra ora tutto il nucleo mimetico-sintetico che la realizzava. Il robot presenta ora, divisi, il suo carattere mimetico, che non mostra più la vicinanza all'umano ma sottolinea la lontananza rispetto a quello a cui vuole somigliare, e l'aspetto sintetico, che si dà come la artificiosità dell'implementazione del momento e del gesto interazionale invece che nella naturalezza tipica del ritmo vivente. In sostanza, è così che si dà il negativo della somiglianza, ovvero si porta in primo piano il divario tra quello che è e quello che dovrebbe essere, un divario che nel tessuto dell'interazione sociale non pareva così incolmabile. L'UVE, dunque, funziona come un meccanismo di capovolgimento dei panorami di senso, in cui le pratiche percettive si riorganizzano e il micromondo di riferimento fa emergere altre salienze, altre pregnanze, che non sono più atte a nutrire l'abito incarnato della reciprocità e dell'interazione sociale ma si dissolvono nella rottura dell'armonia degli effetti. Ora il robot è diventato macchina, mimesi dell'umano come Galatea ancora statua, come il burattino di legno con il sogno di diventare un bambino vero, e anche sua sintesi, orologio condizionato dai meccanismi interni, un congegno del tipo delle macchine cibernetico-computazionali che realizza forme sintetiche di comportamento. Il sincretismo si è frantumato sotto il peso delle aspettative non rispettate, l'intenzionalità guadagnata dalla fiducia che l'interazione aveva riposto nelle capacità della macchina è sbiadita e il robot si esibisce nella sua forma artefattuale, le cui componenti sono ormai venute alla luce. Possiamo descrivere così, d'altra parte, riprendendo gli argomenti del capitolo precedente, il passaggio tra micromondi, ovvero il cambiamento di strategia da *intentional* a *design stance* nella pratica discorsiva per descrivere le azioni del robot: avendo perso l'effetto sincretico, il robot viene percepito principalmente come macchina sintetica e mimetica, capace di mostrare a tratti capacità interazionali ma, di fatto, risultato di meccanismi computazionali che ne permettono il movimento.

Ora, a partire dallo statuto mimetico-sintetico del robot, il modo attraverso cui il micromondo alternativo dell'essere umano in HRI si struttura dipende da tutti quei fattori che abbiamo elencato in precedenza: dall'esperienza e dalla storia interazionale dell'agente, dai *breakdown* già esperiti, dalle

possibilità aperte dall'interazione stessa, dal contesto di realizzazione, dalle forme di vita e dalle espressività pragmatiche disponibili al momento come percorsi di *sense-making* enattivabili. Oltre a questi fattori, il processo tramite cui il micromondo si struttura dipende anche dall'insieme delle possibilità interpretative conosciute, ovvero dai modelli narrativi di cui l'agente si è appropriato nel corso della sua storia di interazioni e dalle occasioni, se presenti, in cui alcune di queste narrazioni sono state impiegate per sciogliere nodi percettivi creatisi in circostanze passate. Come abbiamo visto nel capitolo precedente (§4.3.3), infatti, il modo attraverso cui cerchiamo di riconfigurare il piano percettivo viene influenzato anche dal ruolo che le *folk psychological narrative* hanno nella strutturazione delle *propositional attitude*. In quella sede, ciò ci ha consentito di elaborare una spiegazione circa le modalità incongruenti di esperienza di fronte al robot: implicitamente compreso come agente intenzionale, esplicitamente ricondotto a una macchina. In questo momento, le *propositional attitude* ci permettono di capire quali siano le possibilità di micromondi aperti a seguito dell'UVE e che, spesso, costituiscono il piano esperienziale in cui l'essere umano è immerso al momento della valutazione dell'esperienza, momento che nella maggior parte dei casi viene sondato per verificare la presenza dell'effetto perturbante: infatti, sono rarissimi i casi in cui l'UVE è stata testata attraverso metodi impliciti di valutazione (ad es. Saygin et al. 2012; Urgan et al. 2015, 2018; Cheetham et al. 2011), mentre la metodologia più adottata è, anche in questo caso, quella che fa ricorso a questionari che registrano i giudizi dei soggetti in seguito all'incontro con il robot. Quello che le *propositional attitude* possono farci vedere è che, se il momento dell'UVE costituisce un'esperienza inusuale che deve essere riportata ai livelli di praticabilità del *sense-making* individuale, allora il modo attraverso cui si costituirà il micromondo sarà influenzato anche dalle modalità rese disponibili dalla nostra capacità interpretativa nutrita dalle narrazioni che circolano nella nostra cultura.

Abbiamo visto come le *propositional attitude* entrino in gioco proprio in corrispondenza di forme più complesse di comprensione e interpretazione delle circostanze esperienziali, ovvero quando non sono sufficienti o disponibili modalità dirette o corporee di intendimento. Nel caso dell'UVE, questo fenomeno apre all'incertezza data dal *breakdown* interazionale e convoca altri abiti, non più relativi solamente alla modalità incarnata e in tempo reale di interazione con l'altro, ma utili a comprendere quale sia l'assetto migliore da stabilizzare al fine di continuare a interagire: un processo di concettualizzazione incarnata, quindi, in cui i significati richiamati non sono stipati all'interno dell'architettura funzionale come concetti linguistici e logicamente strutturati, ma hanno la forma di richiami pratici alla sfera interpretativa situata dell'individuo. In questo senso, si comprende anche la differenza interpretativa a livello culturale rilevata negli studi sulle differenze dell'UVE tra popolazioni con *background* interpretativi molto differenti (cfr. §4.3.3).

Pur avendo perso il suo potenziale come agente capace di interazioni sociali, dunque, il robot si configura come un diverso tipo di interlocutore a seconda che le *propositional attitude* dell'individuo che si trova a incontrarlo siano nutrite da narrazioni che dipingono il robot come macchina collaboratrice, amica e solidale (come nel caso, ad esempio, delle storie giapponesi) o come una macchina sterminatrice, ingannatrice e malevola, che vuole imbrogliarci nascondendo la sua vera natura e prendere il nostro posto (come accade nella maggior parte delle narrazioni occidentali). Quindi, il micromondo che si costituisce sarà condizionato sia dalle personali e caratteristiche modalità di interpretare il robot nella cultura di riferimento, sia da tutti gli altri fattori che abbiamo descritto come fondamentali per l'emersione del panorama di senso: abiti sensorimotori, aspettative conseguenti, contesto di interazione, disposizione affettiva, propensione ad affrontare esperienze nuove, tratti della personalità<sup>193</sup>, e così via. Ovviamente, come abbiamo cercato di mostrare nel caso dell'analisi della *humanlikeness* nel dominio della HRI, questi fattori non possono essere considerati universalmente validi: differenze nei diversi stili interazionali o nelle concezioni dei robot possono essere rintracciate senza dubbio anche all'interno dello stesso *background*. La nostra ipotesi, infatti, si costituisce proprio a partire dal fatto che è l'esperienza incarnata a convogliare tutti i fattori che abbiamo elencato nelle pratiche percettive e interpretative dell'individuo, e che ogni circostanza possa richiamare e concatenare abiti in maniera differente. La proposta che avanziamo, quindi, concerne la capacità delle pratiche narrative di costituire il primo appiglio del senso in disarmonia generato dal *breakdown* dell'UVE, un'interpretazione già sedimentata in un abito interpretativo che costituisce un possibile percorso del senso, senza necessariamente essere universalmente valido. Quello che apparirà nel micromondo alternativo non può essere conosciuto in anticipo; dobbiamo attendere che il gioco sviluppi le sue dinamiche.

#### 4.4.10.3 *Esiti in itinere: uncanniness, micromondi enattivi e il robot come monstrum*

Per concludere la nostra analisi dell'UVE, vogliamo sottolineare due riflessioni che conseguono dalla lettura data in questo capitolo sul fenomeno.

Da una parte, gli strumenti concettuali che abbiamo convocato per rendere conto dell'esperienza perturbante di fronte al robot ci hanno permesso di ampliare lo sguardo, fino ad includere fattori fino ad ora lasciati completamente al di fuori dell'indagine. Abbiamo mostrato in più momenti che la corporeità, la capacità di sentire il ritmo interazionale creato dalla commistione degli abiti e la reciprocità dell'interazione sociale siano la chiave per indagare il fenomeno dell'UVE, così oscuro, forse, perché tocca le corde della socialità tra umani. Ora, possiamo anche vedere che tale

---

<sup>193</sup> Un'interessante linea di ricerca sta approfondendo il ruolo dei tratti di personalità all'interno della HRI. Per una panoramica si veda: Walters et al. (2005); Lionel et al. (2020); Gargano et al. (2022).

impostazione epistemologica ci consente di raccogliere le ipotesi esplicative presenti nello stato dell'arte attuale sulle cause dell'UVE e vedere ognuna come declinazioni parziali del nostro approccio. Infatti, le cause addotte dalle spiegazioni percettive, rispettivamente la repulsione data dalla paura per la propria salute (PAH), dalla sensazione di morte (MSH) e dall'estetica evitante del robot (EAH) sono concepibili come alcuni dei significati aperti dalla rottura interazionale dell'UVE, ovvero le prospettive percettive aperte dal micromondo alternativo. In questo caso, da essere considerate cause, nel nostro approccio questi significati diventano effetti pragmatici dell'UVE.

Per quanto riguarda le spiegazioni cognitive, anche queste possono rientrare all'interno del nostro sguardo, se tralasciamo volutamente le incompatibilità epistemologiche e metodologiche e consideriamo il principio alla base dell'UVE. Nella CUH, il momento di incertezza dovuto all'indecisione di fronte al robot è raffrontabile al momento di *breakdown*, ma non nei termini categoriali e computazionali descritti dalla teoria della percezione adottata. L'incertezza c'è, ma non nel decidere lo statuto di cosa ci si trova davanti, processo che secondo l'approccio che abbiamo adottato è realizzato in concomitanza dell'enattivazione dei significati percettivi innescati dagli abiti sensorimotori. Così, possiamo anche sorpassare il cortocircuito creato dalla non sovrapposizione del momento di sforzo cognitivo maggiore (considerato la causa dell'UVE nella CUH) al confine tra categorie concettuali: se vediamo il processo di comprensione non in maniera discreta e puntuale, ma fluida e dinamica, possiamo comprendere che il confine tra categorie non può costituire un elemento significativo né in termini di percezione, né nel fenomeno dell'UVE. Considerando la VEH, invece, possiamo mostrare un accordo riguardo al ruolo delle aspettative nel processo percettivo, ma non nei termini del *predictive coding*. Come abbiamo più volte sottolineato, questa ipotesi ha il merito di introdurre all'interno del dibattito il ruolo del movimento del robot, considerato una causa scatenante in confronto all'apparenza umana dell'androide. In questo senso, non possiamo che trovare dei punti in comune con la VEH, anche se non sono condivise le modalità di spiegazione del fenomeno: la differenza maggiore sta nel fatto che, se la VEH individua il *mismatch* tra movimento meccanico e apparenza biologica dell'androide come elemento chiave per l'emersione dell'UVE, secondo la visione che abbiamo proposto l'isolamento delle caratteristiche non è sufficiente, né necessario – e anzi potenzialmente fuorviante – per la comprensione del fenomeno, oltre che della cognizione in generale. Introducendo il movimento in una prospettiva più ecologica, d'altra parte, riusciamo a comprendere non solo che questo sia un elemento fondamentale da prendere in considerazione per rendere conto dell'UVE, come però lo è la sua assenza. Infatti, la causa principale che abbiamo rintracciato nell'UVE non consiste tanto nella dimostrazione di un movimento non conforme alle aspettative aperte dall'aspetto del robot, quanto invece al fatto che non si muova *abbastanza* e soprattutto con la fluidità che rintracciamo solitamente durante l'interazione sociale. In sostanza,

anche se aspettative, movimento e percezione sono al centro della VEH, il nostro approccio ne ricalibra i significati e inserisce i fattori all'interno di una processualità dal complesso sviluppo spaziale, temporale e sociale. Ancora, in merito alla MPH, il nostro approccio consente di vedere la mente non come un attributo, insieme all'esperienza e all'agentività, da conferire all'altro, ma come qualcosa che si può rintracciare solo nell'incontro situato. Possiamo far valere per la MPH le motivazioni che abbiamo elaborato a favore di una sostituzione del concetto di attribuzione di intenzionalità con quella di riconoscimento, individuazione della stessa: la mente, l'autonomia, l'agentività, l'esperienza sono capacità espresse dall'agente in accoppiamento strutturale con il proprio ambiente, che si realizzano nel movimento vivo e nella socialità dell'interazione in atto. Oltre a questo, possiamo anche notare che, nel caso della percezione del movimento biologico, in cui è chiaro che il robot umanoide viene percepito come un agente sociale, non è stato registrato alcun effetto perturbante e, al contrario, questo aspetto conduceva alla considerazione del robot al pari dell'essere umano. In sostanza, la MPH può essere smontata come causa dell'UVE, come è stato fatto anche altrove (Wang et al. 2017; Wang & Rochat 2020; MacDorman 2024), anche alla luce dei vantaggi reali annoverati negli studi sull'antropomorfismo (cfr. §3.4.1) e a quelli elencati a favore di una sostituzione dei presupposti della ToM (§3.2). Infine, la DH si presenta, probabilmente, come la più vicina all'ipotesi elaborata nel nostro lavoro se consideriamo i termini con cui l'UVE viene descritta. Infatti, la considerazione dell'UVE in termini processuali e non puntuali e il cambiamento dello statuto interazionale del robot costituiscono due punti in comune. Tuttavia, come si può notare dalla descrizione del processo di deumanizzazione, questo non sembra essere una conseguenza dell'UVE, ma il suo motivo scatenante:

Critically, the uncanny replica is not perceived to be a typical robot; it becomes dehumanized “robotic-like” human who lacks humanness. [...] The more human observers attribute humanlike characteristics to (i.e., anthropomorphize) a human replica, the more likely detecting its mechanistic features triggers the dehumanization process that would lead to the uncanny feeling. (Wang et al. 2015: 402)

Anche in questo caso, dunque, ci troviamo a dover operare delle differenze, in quanto crediamo che la rottura dell'aspetto sincretico del robot (quello che può essere considerato il processo di deumanizzazione se visto nei termini della DH) non costituisce la causa dell'UVE, ma un suo effetto. In più, mentre la DH istituisce uno stretto legame con la MPH pur criticandone alcuni aspetti (Wang et al. 2015), l'approccio che abbiamo delineato ci permette di oltrepassare la visione inferenziale posta alla base di entrambe queste ipotesi e considerare sia l'UVE, sia i suoi possibili effetti, come risultati emergenti dalle dinamiche interazionali che chiamano in gioco elementi esterni alla mera computazione cognitiva dell'agente. Ci troviamo, invece, molto più in linea con altri approcci a

supporto della DH, che individuano proprio nell'assenza di una coordinazione motoria l'origine del processo di deumanizzazione:

True, the use of organic-looking materials to realize skin and hair [...] could give the android a very realistic look. But this is only insofar as the android does not move. Once in motion, the lack of the about 600 muscles which compose a human body, of their fluidity and harmony, is immediately detected by a human observer and the uncanny effect appears. (Angelucci et al. 2016: 126)

La seconda riflessione che vogliamo sviluppare concerne proprio il ruolo del movimento all'interno della ipotesi esplicativa che abbiamo costruito. Nell'ipotesi originale dell'*Uncanny Valley Effect*, lo stesso Mori ipotizzava che il robot in movimento acquisisse l'effetto, portando ancora più in profondità il fondo della valle e rendendo la risalita della curva verso la *shinwakan* ancora più ripida. Crediamo che tale intuizione, sebbene anch'essa non provata sperimentalmente ai tempi della formulazione originale, possa acquisire nuovo interesse nell'ipotesi che abbiamo formulato. Infatti, fino a questo momento, negli esperimenti in cui si è cercato di comprendere se il movimento effettivamente accentuasse l'UVE (Thompson et al. 2011; Piwek et al. 2014; McDonnell et al. 2012) i risultati sono stati discordanti. Nel nostro quadro, l'ipotesi sembra essere testata in modi non appropriati: in tutti e tre gli studi citati, infatti, il ruolo del movimento è verificato attraverso variazioni alla cinematica di avatar virtuali, riportando il fallimento alla scarsa *humanlikeness* mostrata dai robot (McDonnell et al. 2012) oppure fornendo prova che l'effetto non cambiasse tra l'osservazione di un'immagine statica e della stessa immagine in movimento (Piwek et al. 2014). Abbiamo visto, però, che una serie di esperimenti effettuati impiegando misure dell'attività neurale testimonia una differenza nelle modalità percettive di robot in movimento da quelli statici (Saygin et al. 2012; Urgen et al. 2013, 2015, 2018; cfr. Vaitonyté et al. 2023) e, anzi, individua nel movimento una delle motivazioni centrali dell'insorgenza dell'UVE. Per alcuni (Vaitonyté et al. 2023: 13), queste evidenze sono almeno parzialmente in conflitto con studi precedenti che evidenziano la capacità dell'essere umano di comprendere i movimenti e i gesti del robot allo stesso modo di quanto accade con quelli di altri esseri umani (Gazzola et al. 2007; Oberman et al. 2007), in un'ottica di mutue esclusioni che vede i meccanismi in modo puntuale, accesi o spenti. Ma se consideriamo gli elementi all'interno del quadro teorico enattivista e situato che abbiamo sviluppato, queste evidenze sono incluse all'interno del flusso dell'interazione incarnata: nel processo di sviluppo dell'incontro con il robot, il fatto che nella struttura cerebrale dell'essere umano si attivino neuroni specchio che rendono possibile la comprensione immediata delle azioni del robot come quelle di un altro partner interazionale non sembra comportare alcun ostacolo teorico o empirico alla rilevazione del fatto che il robot in movimento comporti una maggiore ampiezza della risposta N400, che invece può essere correlata a una suscettibilità della rete neuronale ai cambiamenti contestuali (Urgen et al. 2018). Se riprendiamo

le caratteristiche che delineano le aspettative dell'essere umano di fronte al robot, la causa dell'UVE non si troverà esclusivamente all'interon di meccanismi interni di registrazione di *mismatch* o di sforzi cognitivi di computazione, ma nella distribuzione delle possibilità di mutuo riconoscimento. In altre parole, non solo l'interazione attiva le mie capacità di *motor resonance*, *mirroring*, imitazione e sincronizzazione: anche l'altro deve essere capace di mostrare le stesse capacità. Il punto è che, nonostante le mie capacità di risonanza motoria siano sollecitate dalla percezione attiva del robot come agente sociale, la mancanza delle sue capacità di risuonare sulle molte dimensioni tipiche dell'interazione intersoggettiva, ovvero a livello motorio, affettivo, intenzionale, agentivo, e così via, comporta la ricalibrazione dei rapporti e la messa in crisi del micromondo intersoggettivo. Per far sì che questo accada, c'è bisogno che l'interazione si sviluppi in vivo. L'UVE, così, viene compreso in quanto mancanza del ritmo della reciprocità, e non solamente come il prodotto di una difficoltà individuale.

Ciò che emerge da questo approccio è che la verifica dell'ipotesi del movimento nell'UVE debba essere effettuata con robot *in movimento*, in interazione almeno con il proprio ambiente (Sciutti & Sandini 2017), non con avatar o immagini modificate artificialmente: sembrerà scontato sottolinearlo, ma, come abbiamo più volte evidenziato nel corso di questa tesi, la maggior parte degli studi che indaga l'interazione tra umani e robot non prende in considerazione l'interazione tra gli agenti e, spesso, nemmeno quella dell'agente robotico con l'ambiente in cui è inserito. Il gioco dell'interazione, allora, ci spinge a situare l'indagine della HRI in un contesto dinamico, in cui l'oggetto di studio sia un robot e non una sua riproduzione in CGI, né delle sue rappresentazioni fotografiche. Solo in questo modo sarà possibile comprendere se effettivamente il movimento aumenta l'UVE, la causi o non contribuisca alla sua emersione.

Siamo giunti al momento della chiusura dei giochi: la musica si è fermata, la danza è stata realizzata, l'incontro con l'altro ha dato, in qualche modo, i suoi frutti. Cosa ci rimane, allora, dopo aver testato l'*uncanny valley effect* nel gioco dell'interazione?

L'ultimo aspetto da considerare concerne il risultato della messa alla prova del robot antropomorfo al gioco dell'interazione e riguarda il recupero di un aspetto specifico dell'UVE, che ad oggi sembra essere offuscato a causa delle influenze internaliste allo studio dei fenomeni cognitivi. L'approccio enattivista che abbiamo sviluppato, d'altra parte, ci consente di portare l'effetto di straniamento e indugio di fronte al robot umanoide a confronto su una dimensione più estesa, distribuita nei rapporti intessuti con l'ambiente che abitiamo e manifesta nei significati esperienziali della vita vissuta. Questa idea costituisce il nucleo centrale della concezione dell'*uncanny* come atmosfera proveniente dal dominio della psichiatria (Fuchs 2019). In questo senso, la sensazione di disagio si realizza come l'emersione di un campo di significati e non solamente come una condizione individuale. L'*uncanny*

si posa sulla superficie degli oggetti e dei luoghi, i quali vanno a comporre una sinfonia perturbante che porta a riconfigurare il processo di produzione di senso in accordo alla sensazione di inquietudine, che si manifesta progressivamente prima come una rottura del mondo dei significati abituali e poi con l'emersione di una sensazione di straniamento che mette in allerta. Come sottolineato da Fuchs,

The uncanny thus lie in a particular, viz. fluctuating relationship of foreground and background: the threat does not emerge as such, but is rather anticipated or guessed at through ambiguity, namely through an ambiguity of the foreground. Therefore, the phenomenon often occurs when the field of perception is characterized by undefined, blurred structures, such as twilight, fog or darkness, in which subtlety and ambiguity reside particularly easy. (Fuchs 2019: 103).

Il rapporto tra figura e sfondo messo in risalto dalla descrizione di Fuchs ci riporta alla rideterminazione della somiglianza del robot, non più considerata nella sua vicinanza all'aspetto dell'essere umano, ma sottolineando ciò che ancora non è sufficiente, ciò che ancora mostra il divario. È così che possiamo rileggere la definizione di Freud di *perturbante*, ovvero come il familiare che diventa sconosciuto<sup>194</sup>, l'*Un-heimlich* (Freud 1919: 17-18; Fuchs 2019: 103) in quanto il contrario del conosciuto, dell'agio dato dalla confidenza con il significato. Questa sensazione può essere compresa solamente se inquadrata come il rapporto istituito tra l'individuo e l'ambiente in cui si muove, tra le capacità di dare significato e le possibilità offerte dal micromondo che abita. La ricerca dei meccanismi interni può rendere conto di ciò che accade nell'individuo alla presentazione di questa atmosfera, ma non delle condizioni distribuite che ne provocano l'emersione, né del modo in cui l'individuo trova il modo di attraversarla.

Concepire l'*uncanniness* come atmosfera che connota il passaggio tra il micromondo del familiare e quello dello sconosciuto ci consente, inoltre, di far emergere un altro aspetto dell'UVE, ovvero quello che rintraccia nella sensazione di inquietudine non solo la paura e il disagio, ma anche la sorpresa, la curiosità, lo stupore.

The ambiguity, or the fluctuation, of the situation between familiarity and strangeness encourages a further reaction, namely *fascination*: the uncanny is often experienced with a mixture of terror and curiosity. The flight tendency of fear is opposed to a component of expectant tension, which makes it difficult to break away from the uncanny impression. (Fuchs 2019: 105)

Ecco, allora, che il robot non appare solamente come pericoloso, inquietante, spaventoso, ma anche meraviglioso, straordinario, interessante, e l'UVE può anche essere inteso come un effetto positivo dell'interazione (Cheetham et al. 2014, 2015; Cheetham (Ed.) 2018). Così, il raccordo tra perturbante, robot e UVE si connota di un senso antico, esprimibile con il termine latino *monstrum*:

---

<sup>194</sup> Questo tipo di sensazione è utilizzata anche per descrivere il fenomeno del *Praecox-Gefühl*, ovvero l'esperienza precoce e sottile di straniamento e disconnessione dalla realtà vissuta, registrata soprattutto nei casi di schizofrenia delirante (cfr. Lobaccaro 2022; Fuchs 2019).



originariamente “prodigio”, nella tradizione letteraria la parola acquista il significato di “ciò che esce dal comune”, “l’essere straordinario”, ciò che è orrifico e stupendamente pauroso (cfr. Fortunati 2000). Questa lettura ci consente di considerare il robot non soltanto come ciò che ci pone davanti ad una difficoltà, che chiede di riorganizzare il panorama di azioni e significati nella maniera dell’indecisione e della ritrosia, ma anche come ciò che stimola la curiosità, l’indagine, il senso di meraviglia di fronte alla creatura mostruosa<sup>195</sup>. Il portato affettivo ed esperienziale implicato nel perturbante viene riportato così alla sua originale matrice estetica, che non necessariamente implica la scissione con l’indagine sui processi cognitivi implicati nella comprensione dell’atmosfera – come testimoniato dalla corrente di studi di neuroestetica cognitiva (Gallese 2021; Di Dio & Gallese 2009) – ma nemmeno la riduzione della prima operazione alla seconda. Ciò che si delinea è quindi un piano pragmatico di possibilità di arricchire la conoscenza di ciò che accade concretamente nell’esperienza con il robot, senza la necessità di ridurre tale esperienza a modelli computazionali o a meccanismi cognitivi interni all’architettura funzionale.

In ultima istanza, tramite il ritorno alla prospettiva estetica ed enattiva del perturbante, riusciamo anche a rendere conto del motivo per cui l’UVE si manifesti proprio in concomitanza di incontri con robot altamente somiglianti all’umano, ma non con altre forme di robot dall’apparenza più meccanica o con altre riproduzioni dell’essere umano, sebbene mostrino altissimi gradi di *humanlikeness* come nei dipinti, ma anche perché l’effetto possa verificarsi nel caso di avatar artificiali o di fotografie di robot. Infatti, la prospettiva che abbiamo costruito permette di comprendere che l’UVE non dipende solamente dal modo in cui l’artefatto somiglia all’umano, qualità che permette di raggruppare sotto la stessa categoria tutte le realizzazioni artificiali di copie e doppi, ma anche e soprattutto dalla corporeità dell’artefatto, dalle forme che l’*embodiment* prende nel processo di costruzione dell’aspetto mimetico, con cui è possibile realizzare anche forme di sintesi del comportamento umano. Comprendere che l’UVE nasce nel contesto dell’interazione *embodied* permette di operare una distinzione all’interno della categoria degli artefatti che assomigliano agli umani, individuando come *uncanny* quelli che enattivano gli abiti incarnati tipici dell’interazione sociale tra umani e da cui ci aspettiamo possano rispondere, tramite azioni, all’apertura del micromondo interazionale che ne deriva. Robot che mostrano maggiori componenti meccaniche non comportano l’enattivazione della serie di abiti incarnati radicati nell’incontro intersoggettivo: le pratiche percettive vengono indirizzate verso altre possibilità di significazione, che non necessariamente presuppongono la mutualità del riconoscimento reciproco. D’altra parte, questo accade anche con i dipinti, a volte anche

---

<sup>195</sup> Non riesco a trovare modi migliori per descrivere l’esperienza delle sensazioni miste ed eterogenee di eccitazione e timore, di euforia e cautela, che ho provato quando, per la prima volta, mi sono trovata di fronte ad Abel (Cominelli et al. 2019): un robot androide dotato della sola parte superiore del corpo che raffigura le sembianze di un bambino di circa 12 anni. La sensazione può essere descritta come l’istinto di avvicinarsi per guardare meglio il prodigio della tecnica che si ha di fronte, ma allo stesso tempo segnata dal timore incarnato dell’inaspettato.

più realistici di certi artefatti materiali, che mostrano individui somiglianti all'umano ma non per questo suscitano effetti perturbanti: in sintesi, non enattivano il micromondo della socialità. Allo stesso modo, la possibilità che l'effetto sia generato tramite l'osservazione di fotografie che rappresentano robot umanoidi o immagini di avatar virtuali è resa possibile per lo stesso motivo: dal momento che percepire è enattivare significati pragmatici in corrispondenza di *affordance* ambientali e sociali, vedere il volto del robot in foto convoca allo stesso modo, ma probabilmente in misura meno rigida<sup>196</sup>, le possibilità interazionali implicate nel riconoscimento intersoggettivo. L'UVE, in questi casi, non è ovviamente il risultato di un'interazione concreta, ma della familiarità con cui i significati percettivi vengono consumati in interazione con il proprio ambiente: insieme agli occhi del robot, percepiamo la possibilità della coordinazione reciproca; insieme alla sua espressione, intuiamo il suo valore affettivo. In questo senso, la dimensione virtuale e immaginativa della cognizione umana, anch'essa basata sulle capacità sensorimotorie e sulla storia dell'individuo (Hutto 2015), configura le stesse predizioni basate sugli abiti delle interazioni incarnate e garantisce una comprensione del significato sulla base della ri-enattivazione di aspettative e pratiche realizzate in interazione con l'altro<sup>197</sup>.

Ci troviamo, quindi, a supportare l'ipotesi che l'UVE sia dipendente dalla qualità estetica esibita dall'artefatto, ma crediamo che il legame fra i due fenomeni sia meglio compreso se mediato dalle circostanze in cui l'artefatto e l'umano si incontrano e dalle modalità di *embodiment* concretamente realizzate, ovvero nella creazione di un'atmosfera interazionale ed esperienziale che coinvolge corpo, ambiente, reciprocità, significati virtuali insiti nelle aspettative e abiti sensorimotori pronti a instradare il senso all'interno del micromondo del presente come momento concreto della realizzazione del *sense-making* (Varela 1992b). In sostanza, una definizione dell'UVE che possa

---

<sup>196</sup> Questa ipotesi si basa sul fatto che, se il movimento acuisce l'UVE, allora l'assenza di movimento dovrebbe allievarne gli effetti. Una delle motivazioni potrebbe essere ricondotta alla minor rigidità delle aspettative coinvolte nella comprensione dell'immagine. Sarebbe interessante, in futuro, testare questa ipotesi sperimentalmente all'interno di un approccio enattivista alla HRI e all'UVE come quello che abbiamo delineato in questa tesi.

<sup>197</sup> Una possibile obiezione potrebbe considerare il fatto che le fotografie che ritraggono esseri umani non comportano lo stesso effetto perturbante. Crediamo che questa obiezione possa essere spiegata considerando i livelli di realtà enattivati dalla pratica (cfr. Eco 1997): anche qui, il contesto gioca un ruolo fondamentale e si concatena, soprattutto, alla storia dell'essere umano (cfr. Bacaro *forthcoming*). L'epoca storica in cui ci troviamo e il livello di avanzamento tecnologico disponibile nella parte più ricca del mondo hanno consentito alla maggior parte degli individui cresciuti in occidente dopo la fine della Seconda Guerra Mondiale di sviluppare una certa confidenza con artefatti fotografici che ritraggono altri esseri umani. In questo senso, quando ci troviamo di fronte a una fotografia, non crediamo che il volto che vediamo al suo interno sia una persona in carne ed ossa, quanto invece una sua rappresentazione. Vedere fotografie di esseri umani è qualcosa che ci è familiare dalla nascita, a differenza di quanto succede con avatar e robot o come accade con altre forme di fotografia sviluppate negli ultimi anni tramite un certo tipo di tecnologia che permette la modifica della rappresentazione quasi in tempo reale. Si pensi, ad esempio, agli studi che stanno cercando di comprendere gli effetti dei filtri Instagram (Isakowitch 2022; Gallese 2024) sulla sfera cognitivo-percettiva degli individui, o all'effetto prodotto da un'immagine palesemente modificata. Anche in questi casi, il rapporto tra realtà e rappresentazione è la chiave per comprendere i fenomeni cognitivi ad essi associati, ma solamente se inseriti all'interno del contesto delle pratiche incarnate.

risuonare con i presupposti che abbiamo individuato è già stata fornita da MacDorman, uno dei maggiori studiosi del fenomeno, che in un'intervista ha descritto così il fenomeno del perturbante:

I think it is that feeling of being in the presence of another human being – the moment when you feel in synchrony with someone other than yourself and experience a 'meeting of minds'. Negative 'shinwakan', the uncanny, is when that sense of synchrony falls apart, the moment you discover that the one you thought was your soul mate was nothing more than smoke and mirrors. (MacDorman in Hsu 2012)

In sintesi, la lettura che abbiamo fornito della HRI e dei fenomeni studiati all'interno di questo campo disciplinare eterogeneo ci porta ad aprire un percorso di ricerca in cui è l'interazione, per prima, ad essere considerata fondamentale, sia per lo sviluppo delle capacità di coordinazione reciproca, sia per la possibilità che si dà solo al suo interno di poter recuperare le rotture insite nel processo di costruzione di senso, sia per arrivare a comprendere quali siano gli aspetti da indagare nella ricerca sul rapporto tra umani e robot. L'interazione si pone in primo piano non solamente come la somma delle singole unità dialogiche, ma come un anello interazionale in cui gli agenti costruiscono un rapporto di mutui concatenamenti nell'incontro tra *embodiment* organismici e organismoidi.

L'ultima operazione che vogliamo svolgere, allora, è quella di iniziare a sondare quali siano alcune delle possibilità aperte da questa riconfigurazione del piano teorico su quello empirico. Nella nostra visione, questo si declina in almeno due conseguenze sulla metodologia di indagine della HRI: da una parte, la possibilità di controbilanciare l'attenzione riposta sui meccanismi interni di funzionamento della cognizione in favore delle dinamiche distribuite che l'agente intesse con il suo ambiente e con gli altri sistemi che lo compongono; dall'altra, il contributo relazionale che il robot può esprimere in interazione e, a fronte della messa in risalto del ruolo del *breakdown* interazionale e della coordinazione inter-agentiva nei processi di reciproco intendimento, la sua capacità di individuare e accomodare in qualche modo le crepe mostrate dalle sue impossibilità di comprensione genuina. Secondo questa visione, la possibilità di spostare l'attenzione dei processi di realizzazione della HRI dai meccanismi cognitivi interni, considerati l'aspetto importante per entrambi gli agenti coinvolti – umano e umanoide – alle capacità esibite in interazione e nel comportamento concede l'opportunità di indagare aspetti non ancora considerati e, allo stesso tempo, ampliare la conoscenza che riguarda le nostre capacità cognitive e interazionali. Apriamo la strada, dunque, alla costruzione di una *Galatea Futura*.

#### 5.4 Galatea Futura

Nel capitolo §3, abbiamo considerato il mito di Galatea come la metafora attraverso cui leggere le possibilità esibite dai robot sociali costruiti per interagire con noi esseri umani: abbiamo quindi

fornito l'analisi della *Galatea moderna*, individuando le abilità di costruire un piano di interazione pragmatica con l'essere umano nel suo aspetto sincretico, armonia degli effetti realizzata grazie al contributo sostanziale dell'interazione con agenti sociali come gli esseri umani. In questo momento, i presupposti epistemologici che abbiamo delineato ci conducono a tradurre l'impianto teorico che tiene insieme il sincretismo del robot e i diversi aspetti delle capacità cognitive umane sul piano pragmatico dell'indagine scientifica. Quello che viene reso possibile, infatti, è riorganizzare gli elementi che abbiamo individuato nella nostra indagine a comporre un piano di possibilità concrete di ricerca, basate anche sulle attuali tecniche di sviluppo dei robot sociali, con il fine di instradare la costruzione di una *Galatea futura*, ovvero robot più capaci di comprendere non solo le dinamiche cognitive che permettono l'attraversamento del micromondo individuale, ma soprattutto di intercettare possibilità concrete di concatenamento e coordinazione con l'essere umano. Per avviare questo percorso, vogliamo iniziare con una possibile riconsiderazione del metodo sintetico, da una prospettiva interazionale ed enattivista.

#### 4.4.11 Metodo Sintetico Enattivista

Nel primo capitolo, all'interno della ricostruzione critica del concetto di macchina che ha portato alla nascita del robot e della robotica, abbiamo chiamato in causa il contributo epistemologico del metodo sintetico alla strutturazione di una concezione particolare e innovativa della macchina. Questo specifico concetto filosofico ed epistemologico di macchina, a partire dalle considerazioni cibernetiche sulla possibilità di comprendere la struttura del pensiero umano tramite la costruzione di artefatti che funzionassero allo stesso modo, ha assunto un doppio ruolo nel panorama delle scienze dell'artificiale: da una parte, modello come qualcosa che astrae il funzionamento dell'oggetto sottoposto ad indagine; dall'altra, modello come qualcosa che serve da guida, ovvero che, se preso ad esempio, indirizza gli sviluppi della teoria e fornisce i vincoli della sua evoluzione (cfr. §1.2.3; Dupuy 1994). In questo doppio terreno costruito dalle nozioni cibernetiche di modello e di macchina, il metodo sintetico si realizza come il movimento elicoidale che procede attraversando fasi di osservazioni del comportamento del sistema *target*, di modellizzazione dello stesso, di implementazione del modello in una macchina, di osservazione del comportamento esibito dalla macchina, di confronto tra i due sistemi – teorico e meccanico – e di ritorno sul sistema artificiale, al fine di ottimizzarne la realizzazione e la performatività (Cordeschi 1998, 2002).

Nel corso degli sviluppi a seguito della costituzione della ricerca sull'intelligenza artificiale, i presupposti di questa metodologia di indagine sono confluiti anche nella formulazione dei principi del funzionalismo cognitivista e nelle speranze che la macchina potesse riuscire a dimostrare capacità cognitive genuine, al pari dell'essere umano. Sono questi i principi posti anche alla base della

formulazione del test di Turing, ovvero quelli che hanno permesso a uno dei primi informatici della storia di poter chiedere «Can machines think?» (Turing 1950: 433). Nel nostro percorso di sviluppo del gioco dell'interazione, abbiamo più volte stabilito un parallelo con il test dell'imitazione e abbiamo rintracciato non solo alcune somiglianze, ma anche molte differenze (cfr. §5.1, 5.3.3.1). L'ultima che vogliamo portare alla luce riguarda proprio la riconfigurazione del metodo sintetico alla luce della struttura partecipativa ed *embodied* del gioco dell'interazione, che include le prospettive enattiviste sulla mente e sulla cognizione che si sono sviluppate a partire da una successiva ed ulteriore evoluzione della cibernetica. Se riprendiamo i presupposti descritti e provati in questo capitolo, infatti, ciò che si prospetta è la possibilità di ampliare la portata del metodo sintetico, ponendo l'accento non tanto sulle possibilità di costruire una macchina che funzioni esattamente implementando i nostri stessi processi cognitivi, ma che esibisca un comportamento *coordinato* rispetto a quello che noi esibiamo in interazione e che ci aspettiamo che il nostro partner sappia individuare e comprendere al fine di parteciparvi: un comportamento *mimetico* nei molti modi in cui può darsi la coordinazione in interazione sociale.

Possiamo affrontare questa rilettura in due punti:

- Da una parte, anche se il metodo sintetico ha condotto alle formulazioni teoriche che hanno caratterizzato il punto di vista funzionalista sulla cognizione e sull'intelligenza artificiale, le sue capacità euristiche hanno nutrito anche la seconda fase del movimento teorico cibernetico, che ne ha colto i presupposti e li ha rielaborati nella costruzione di artefatti basati su intuizioni differenti rispetto a quelle poste alla base dell'impresa dell'AI. Si ricorderà, infatti, che l'operazione effettuata da von Foerster (1987, 2003) alla costituzione del BCL aveva come intento quello di riconfigurare le premesse cibernetiche confluite nella costruzione dei sistemi esperti e dei risolutori di problemi che hanno contraddistinto lo sviluppo dell'AI negli anni '60 e '70, a partire da una differente visione sulla natura della cognizione. In quel periodo, lo sviluppo di concetti come quello di *auto-organizzazione* prima (von Foerster 1960) e di teorie come quella dell'*autopoiesi* poi (Maturana & Varela 1972, 1980) ha portato alla riconfigurazione sia delle idee su cognizione e intelligenza che sarebbero poi confluite nel lavoro di Rodney Brooks e nella *nouvelle AI* (cfr. §2.2.2), sia di alcune modalità di pensare all'artificiale in relazione all'umano. Queste nuove modalità si concentravano sull'importanza della chiusura organizzazionale dei sistemi naturali complessi, mettendo in risalto il fatto che il ruolo del cervello e delle connessioni neurali può essere compreso pienamente solo se si riconosce come parte di un complesso organizzazionale più ampio, che comprende anche la struttura materiale attraverso cui le capacità cognitive sono realizzate e l'ambiente all'interno nel quale tali processi si concretizzano (cfr. Scott 2004). Esempi come l'automa cellulare

Bittorio, costruito da Varela (1988; Varela et al. 1991) per modellizzare la chiusura organizzazionale degli esseri viventi e il ruolo delle perturbazioni ambientali nella costituzione del *sense-making* (cfr. Beer 2019), sono i primi risultati espliciti dell'incontro tra enattivismo e cibernetica di secondo ordine, e tuttavia riprendevano una linea di ricerca già sviluppata nella prima fase del pensiero cibernetico attraverso la realizzazione di artefatti capaci di mostrare un grado di accomodamento con le perturbazioni provenienti dall'ambiente<sup>198</sup>. Macchine come l'omeostato di Ashby e le tartarughe costruite da Grey Walter (cfr. §1.2.3) mettevano in primo piano il rapporto tra l'ambiente e i processi di comportamento del sistema osservato. In questo senso, il metodo sintetico, nella sua doppia accezione cibernetica, prevede costitutivamente la possibilità di guardare non solo ai meccanismi interni di cognizione, ma anche al rapporto instaurato dall'organismo con il proprio ambiente. Seguendo la rilettura che abbiamo proposto, allora, troviamo il modo di rinnovare il proposito cibernetico della messa alla prova della macchina nel contesto relazionale con l'ambiente, inserendo all'interno del quadro interazionale due variazioni: da una parte, la considerazione che il processo stesso si costituisca in accordo con le occasioni e con i vincoli rilevati nell'ambiente e, dall'altra, l'introduzione all'interno dell'ambiente di un insieme di *affordance* particolari, ovvero quelli che identificano un altro agente con possibilità di interagire socialmente.

- Lo sguardo alla modellizzazione, dunque, si sposta dai meccanismi interni all'agente alle possibilità interazionali implicate nei rapporti con l'ambiente e i suoi richiami all'azione, che nel caso della HRI comprendono anche le possibilità sociali di coordinazione reciproca. In questo senso, il metodo sintetico riesce a raccogliere le spinte enattiviste all'esternalizzazione dei processi cognitivi e a includere nella sua metodologia la possibilità di partire dall'interazione per modellizzare il contributo dei singoli, e non solamente il contrario. Nel metodo sintetico cibernetico, infatti, l'interazione adeguata con l'ambiente, ovvero quella che rispecchiava le modalità di comportamento esibite dal sistema *target*, veniva interpretato come la conferma delle ipotesi di modellizzazione e implementazione che avevano guidato la costruzione della macchina: se il comportamento esibito era conforme a quello individuato nel sistema *target*, questo significava che l'ipotesi era ben sviluppata e che gli stessi meccanismi cognitivi potessero essere realizzati su più supporti e attraverso diverse modalità di implementazione. Se spostiamo lo sguardo alle idee enattiviste sull'interazione che abbiamo esposto nel corso della tesi, d'altra parte, riusciamo a realizzare l'operazione inversa: a partire

---

<sup>198</sup> Oltre a confluire nelle idee di Brooks per la *situated robotics*, un ulteriore sviluppo di questa corrente di pensiero consiste nello studio dei sistemi cognitivi come sistemi dal comportamento adattivo, presupposto alla base della costruzione di artefatti artificiali con capacità di adattamento alle perturbazioni dell'ambiente tramite evoluzioni di algoritmi e reti neurali (Beer 1990, 1997).

dall'osservazione delle dinamiche interazionali mostrate dal sistema *target* in interazione con il proprio ambiente, che nel nostro caso sono costituite dalle dinamiche sociali di coordinazione sensorimotoria e semiotica nell'interazione tra umani, allora possiamo ipotizzare modelli sintetici di interazione e implementarli nelle architetture cognitive di macchine artificiali sulla base di ciò che l'interazione stessa chiede ai singoli. È un processo di ribaltamento delle prospettive, che si pone in linea con la considerazione del comportamento degli agenti e del loro stesso potenziale interazionale come effetto dell'incontro reciproco, e non come causa. Se è l'interazione sociale a costituire ognuno degli agenti come interattante (De Jaegher et al. 2010; Di Paolo et al. 2007), allora per costruire macchine capaci di occupare le posizioni lasciate aperte dall'interazione si può operare un'inversione dello sguardo e partire proprio da qui per modellizzare le capacità interazionali dei robot in HRI.

In sostanza, il cuore di un *metodo sintetico enattivista* consiste nel considerare prima di tutto il potenziale sociale aperto dall'incontro fra robot e umani come sistema *target*, ovvero come il comportamento da sintetizzare. I meccanismi cognitivi interni al sistema, che nella versione cibernetica classica del metodo costituiscono il punto attentivo, lasciano il posto alle dinamiche dell'interazione, diventando queste ultime il fulcro dell'indagine. Le conseguenze relative alla composizione interna dei principi organizzazionali non derivano direttamente dalla possibilità di realizzare forme sintetiche di coordinazione nel comportamento sociale, ma dalla capacità del sistema di mantenere la propria autonomia interazionale e la socialità del sistema-interazione nel suo complesso. Questo si traduce in una doppia condizione di verificabilità: da una parte, il sistema *target* è modellizzato se l'essere umano riesce a intrattenere interazioni sociali con agenti diversi da quelli biologici, sviluppando forme di concatenamento sociale che usualmente si realizzano proprio con altri agenti biologici: ciò significa portare a termine dei compiti in cui è necessaria una collaborazione, ma anche affidarsi alle capacità sensorimotorie della macchina, ad esempio nei contesti di cura e assistenza (Broekens et al. 2008; Parviainen et al 2018); dall'altra, per la macchina significa riuscire a cogliere le strutture del ritmo nell'incontro con l'altro e sfruttare le rotture per evolvere la propria conoscenza dei fenomeni di aggiustamento che avvengono con l'altro. Il primo di questi due punti è considerabile uno degli obiettivi a lungo termine della robotica sociale, sviluppato tramite i molti tentativi di realizzare forme accettabili di comunicazione e collaborazione. D'altra parte, la seconda condizione di verificabilità si configura come la possibilità di una ricerca all'interno della tecnologia robotica che si costituisce allo scopo di *insegnare* al robot a mostrare comportamenti adeguati in contesti interazionali che ne richiamano lo svolgimento. Nel prossimo paragrafo, cercheremo di affrontare questo ultimo aspetto analizzando come, a partire da tecnologie già esistenti, sia possibile

ricalibrare l'indagine includendo i presupposti enattivisti sulla socialità, sull'*embodiment* e sui processi predittivi in cui si intrecciano abiti incarnati e aspettative di coordinazione.

#### 4.4.12 *Kintsugi per l'HRI, o l'arte di tenere insieme i pezzi (dell'interazione): il caso della Learning from Demonstration*

Al fine di comprendere come sia possibile dotare il robot della capacità di concatenarsi agli abiti sensorimotori dell'essere umano e occupare quelle posizioni aperte per lui dall'interazione, proviamo a pensare al modo attraverso cui noi, esseri umani, arriviamo a poter esercitare sapientemente questa abilità, ovvero ai processi di apprendimento.

Come abbiamo visto in precedenza, gli esseri umani imparano a gestire i propri abiti sensorimotori, le proprie aspettative e a riconoscere le possibilità di concatenamento tramite un processo di costruzione di una storia interazionale attiva in cui, fin da bambini, si è immersi in un sistema sociale e culturale che richiede il nostro contributo: così, i *caregiver* allestiscono posizioni conversazionali e dialogiche incarnate che il bambino si trova ad occupare, consentendo quindi al bambino di acquisire progressivamente le capacità di abitare abilmente quelle posizioni lasciate vuote per lui e maturare la possibilità di presentarne a sua volta. Costruire robot capaci di apprendere allo stesso modo è uno dei sogni della robotica già a partire dagli anni '60, ed è stato rinforzato nel corso del tempo dalle frontiere raggiunte nel campo dell'apprendimento artificiale come il *reinforcement learning*, le architetture *behavior-based* di Brooks (§2.3.2.1), gli algoritmi genetici elaborati dalla robotica evolutiva, la *developmental robotics* (Cangelosi & Schlesinger 2015; Lungarella et al. 2003), le reti neurali (cfr. Chan et al. 2012) e, molto di recente, i modelli di intelligenza artificiale generativa (cfr. Katara et al. 2023; Yuan et al. 2023; Wang et al. 2025). Una delle tecniche più indagate, e che sembra presentare ottime premesse dati i risultati raggiunti fin qui, è quella dell'apprendimento tramite dimostrazione, o *Learning from Demonstration* (LfD; Argall et al. 2009; Mohammad & Nishida 2015; Lee 2017; Calinon 2018; Ravichandar et al. 2020).

La LfD è quel paradigma di apprendimento automatico secondo cui il robot acquista nuove abilità apprendendo come imitare il comportamento realizzato direttamente da un dimostratore umano (Argall et al. 2009; Ravichandar et al. 2020). Quando è stato introdotto, intorno alla fine degli anni '90, veniva considerato il modo attraverso cui si potesse raggiungere la capacità di costruire robot umanoidi con capacità molto simili a quelle umane (Schaal 1999), fiducia che ha innescato anche indagini nel campo delle scienze cognitive e della HRI al fine di comprendere quali fossero i meccanismi alla base delle capacità di apprendimento per imitazione esibite dagli esseri umani e da alcuni animali sociali (Dautenhahn & Nehaniv (Eds.) 2002). Questo tipo di indagine ha contribuito in modo sostanziale a definire i problemi e a intradare i possibili percorsi di risoluzione nel campo



della LfD. In particolare, Nel volume curato da Kerstin Dautenhahn e Christopher Nehaniv, i due autori propongono la formulazione di un problema che, ancora oggi, si ritrova al cuore delle tecniche di apprendimento automatico basati sull'osservazione, ovvero il problema della corrispondenza (*correspondence problem*, Nehaniv & Dautenhahn 2002).

Il problema della corrispondenza viene identificato a partire da un'osservazione:

The identification of any form of social learning, imitation, copying, or mimicry presupposes a notion of correspondence between two autonomous agents. Judging whether a behavior has been transmitted socially requires the observer to identify a mapping between the demonstrator and the imitator. [...] To achieve a behavioral match, of whatever sort, a correspondence must explicitly or implicitly be present. In trying to imitate or learning socially by observing, an autonomous agent must somehow solve the correspondence problem. (Nehaniv & Dautenhahn 2002: 41, 43)

In sintesi, la mimesi è concepita come la capacità del sistema di estrarre la struttura del comportamento da replicare e di tradurre tale comprensione nella dimostrazione dello stesso pattern di movimenti, mostrando solo così di aver risolto il problema della corrispondenza, definito formalmente in questi termini<sup>199</sup>:

Given an observed behavior of the model, which, from a given starting state, leads the model through a sequence (or hierarchy) of subgoals (in states, action, and/or effects, while possibly responding to sensory stimuli and external events), find and execute a sequence of actions using one's own (possibly dissimilar) embodiment, which, from a corresponding starting state, leads through corresponding subgoals (in corresponding states, actions, and/or effects, while possibly responding to corresponding events). (Nehaniv & Dautenhahn 2002: 43)

L'esempio più utilizzato per illustrare la complessità del problema è quello della corrispondenza tra le capacità di volo mostrate dagli uccelli e quelle artificialmente realizzate degli elicotteri. Nonostante le differenze strutturali tra questi due sistemi, infatti, è possibile stabilire una corrispondenza funzionale tra stati e movimenti per replicare alcune caratteristiche di volo: negli uccelli, l'angolo delle ali determina la direzione e l'intensità del movimento, mentre negli elicotteri questa funzione è svolta dall'angolo del rotore. Per simulare il comportamento di *hovering*, ossia la capacità di rimanere sospesi in aria mantenendo una posizione fissa, le ali degli uccelli e il rotore dell'elicottero si allineano orizzontalmente, bilanciando perfettamente la spinta verso l'alto con il peso. Quando questo angolo si inclina rispetto all'orizzonte, il cambiamento genera una componente di spinta orizzontale che consente il movimento in avanti o all'indietro. In sostanza, ci troviamo di fronte a una variazione del principio funzionalista della realizzabilità multipla, secondo cui è possibile realizzare una funzione (cognitiva o, come in questo caso, motoria) individuando le caratteristiche

---

<sup>199</sup> In termini meno formali, il problema è esposto come una domanda in un'altra sezione del volume: «given an imitator (a biological or artificial system) trying to imitate a model (the biological or artificial system to be imitated), how can the imitator identify, generate, and evaluate appropriate mappings (perceptual, behavioral, cognitive) between its own behavior and the behavior of the model?». (Dautenhahn & Nehaniv 2002: 1)

del pattern funzionale e riconvertendolo per essere realizzato su un supporto differente. Le capacità di mimesi, inoltre, ritenute essere al centro della capacità di comprensione del sistema, vengono modellizzate come operazioni di astrazione inferenziale strettamente legate, nei contesti sociali, alle capacità di *mindreading* e *Theory of Mind*: «Understanding and matching higher-level structures characterizes various other forms of observational learning and matching involving theory of mind or mind-reading» (Nehaniv & Dautenhahn 2002: 53; cfr. Mitchell 2002; Demiris & Hayes 2003).

Lo schema delineato dal problema della corrispondenza viene ripreso esattamente nella stessa forma all'interno delle ricerche sulla LfD (Argall et al. 2009; Lee 2017; Ravichandar et al. 2020). In particolare, seguendo gli sviluppi della tecnologia in questo campo e inglobando riflessioni provenienti dal dominio delle scienze cognitive, il problema nella LfD viene sottoposto anche a ulteriori elaborazioni, allo scopo di scomporre il problema in sottounità più dettagliate. In particolare, Mohammad & Nishida (2015: 25) forniscono una lista completa dei problemi rintracciabili nella LfD, che devono essere affrontati se si vuole dotare il robot delle capacità di riprodurre il pattern di azioni dimostrato dall'essere umano:

- *Action segmentation*: Questo compito consiste nel riconoscere i confini tra i diversi comportamenti elementari all'interno del flusso continuo di azioni esibite dal dimostratore. Questo processo viene considerato essenziale per scomporre l'azione complessa in unità discrete che il robot può comprendere e replicare. Ad esempio, osservando un umano che prepara una tazza di tè, è utile segmentare il flusso di azioni in step come “prendere la tazza”, “riempire il bollitore”, e “versare l'acqua”, ciascuno dei quali rappresenta un confine tra azioni distinte e significative per la replica.
- *Behavior significance for imitation*: Si tratta di identificare quali comportamenti del dimostratore possano risultare interessanti o rilevanti per essere imitati, ovvero individuare le caratteristiche essenziali che un robot dovrebbe replicare. Alcuni movimenti potrebbero non essere necessari per la riproduzione del comportamento interessato (ad esempio, il modo in cui ci si lega i capelli prima di fare un panino), mentre altri, come la presa sicura di un oggetto fragile o pericoloso (ad esempio, un coltello con cui tagliare il pane per il panino), devono essere riprodotti con precisione per garantire l'efficacia dell'azione o la sicurezza del contesto. Quindi, un robot che imita un umano dovrebbe poter decidere di ignorare il movimento di legatura dei capelli ma di replicare fedelmente la presa del coltello.
- *Perspective taking*: Questo aspetto consiste nel comprendere la situazione dal punto di vista del dimostratore, ovvero il fatto che il robot possa riuscire ad immaginare la prospettiva che il dimostratore ha nel contesto, utilizzando i propri sensori. Questo implica che il robot tenti di interpretare il flusso di informazioni sensoriali come farebbe l'umano. Ad esempio, se l'azione

richiede di afferrare il coltello appoggiato sul tavolo, il robot potrebbe dover considerare come l'oggetto appare dal punto di vista del dimostratore, adattando la propria percezione per replicare l'azione.

- *Demonstrator modeling*: Il compito richiede di capire quali siano i “primitivi d'azione” specifici che il dimostratore sta utilizzando per realizzare il comportamento. Questo significa analizzare la sequenza di movimenti e individuare la loro relazione con gli input sensoriali che guidano tali azioni. Per esempio, un dimostratore che afferra una palla potrebbe attivare specifici movimenti della mano in risposta alla percezione visiva della palla stessa. Il robot dovrà quindi comprendere questa connessione per poter riprodurre con successo l'azione.
- *Correspondance problem*: Come da definizione, si tratta di affrontare la sfida di mappare i movimenti e le azioni del dimostratore sul corpo dell'osservatore, che può avere una struttura fisica differente, e nel contesto di riferimento. Ad esempio, il robot che deve imitare un umano che prepara un panino deve tenere conto dei gradi di libertà dei suoi bracci robotici e metterli in relazione al movimento realizzato dalle articolazioni delle braccia umane, e dovrà trovare una corrispondenza che mantenga il significato dell'azione pur adattandola alla propria struttura.
- *Evaluation problem*: Consiste nel determinare come l'osservatore possa verificare se l'imitazione è avvenuta correttamente e misurare la qualità dell'imitazione per perfezionarla. Ad esempio, se un robot imita il movimento di versare dell'acqua in un bicchiere, l'efficacia dell'imitazione potrebbe essere valutata dal livello di precisione nel versare il liquido senza fuoriuscite. Un *feedback* utile potrebbe provenire sia dal risultato dell'azione, ovvero dalla constatazione che il bicchiere sia stato riempito, ma anche dalla validazione esplicita del costruttore.
- *Quality improvement problem*: Infine, in questo caso si tratta di esplorare come l'osservatore possa migliorare progressivamente la qualità dell'imitazione nel tempo, sia adattandosi a nuove circostanze, sia modificando i movimenti appresi per rappresentare meglio l'intenzione del dimostratore. Ad esempio, se un robot impara inizialmente a sollevare una scatola con entrambe le mani, potrebbe adattarsi e riuscire a sollevarla con una sola mano in contesti dove lo spazio è ristretto, ottimizzando così l'efficacia e la naturalezza dell'azione imitata.

Come notato dagli stessi autori, le questioni maggiormente affrontate sono relative al *perspective taking*, al problema del dimostratore e a quello della corrispondenza (cfr. anche Argall et al. 2009; Lee 2017). In sostanza, le principali preoccupazioni relative alla capacità di implementazione della LfD riguardano la capacità di assumere la prospettiva percettiva dell'altro, la capacità di scomporre

l'azione in sequenze elementari che portano alla realizzazione del comportamento e comprendere come sia possibile, a partire dalle unità gestuali disponibili, realizzare una copia dell'azione osservata, mettendo in pratica così la propria capacità imitativa. Il robot, che ha lo scopo di replicare nel modo più accurato possibile il comportamento esibito dal dimostratore, deve operare una scomposizione del flusso di azioni, mapparne il perimetro e riportare la cinematica alle sue possibilità di esecuzione, cercando nei propri modelli la possibilità di replicare l'azione osservata e mappata.

Ora, a fronte delle considerazioni che abbiamo riportato nel corso della nostra trattazione, possiamo senza riserve ricondurre questo tipo di concettualizzazione della comprensione dell'azione e delle facoltà imitative a un preciso paradigma teorico: quello della ToM. Come già anticipato nel caso dei primi studi sull'imitazione propedeutici allo sviluppo artificiale, il modello teorico di riferimento che viene assunto come base per spiegare i meccanismi di comprensione reciproca è proprio quello che vede alla base della cognizione sociale l'utilizzo di abilità logico-inferenziali, attraverso cui il sistema cognitivo verrebbe a conoscenza delle intenzioni dell'altro tramite ragionamenti sui suoi stati mentali (cfr. 3.2.4). Questo approccio è esplicitamente adottato anche nella LfD:

Humans understand the behavior of other humans (and some animals) by assigning mental states to them and reasoning about how these states affect or generate their perceived behavior. This ability to reason about the mental state of other agents (whether people or otherwise) is called by various names by different researchers including mentalizing, mind reading, and theory of mind (ToM). (Mohammad & Nishida 2015: 211)

Questo tipo di concettualizzazione ha come conseguenza una serie di posizioni metodologiche che riguardano la modellizzazione dell'azione dell'altro a fini imitativi:

- i) In primo luogo, l'azione è vista come un flusso discretizzabile di azioni elementari, la cui combinazione e sequenza dà vita al comportamento. L'idea che l'azione possa essere vista come un flusso di unità discrete o "elementi di base" si basa su una segmentazione artificiale, tipica degli approcci computazionali tradizionali, non troppo lontana da quella che ha informato la costruzione di robot come SHAKEY (cfr. 2.1.2), anche se con tecnologie molto più avanzate a disposizione.
- ii) Nella LfD basata sulla ToM, il processo imitativo passa per una comprensione logico-linguistica degli scopi e delle intenzioni del dimostratore, il cui comportamento deve essere scomposto e computato al fine di essere mappato e convertito nelle possibilità di esecuzione dell'osservatore. Tale impostazione proviene dall'idea secondo cui «*Imitation of action requires the representation of goals as well as motor mimicry and hence could be a way for the subject to become aware of another's goal, an important step on the path to mind reading*» (Gerrans 2009: 911).

- iii) Abbiamo visto come, nella maggior parte dei casi, abbracciare la ToM significa adottare una delle sue versioni più diffuse: la TT o la ST. Entrambi questi modelli ipotizzano l'uso di rappresentazioni mentali nel processo di comprensione dell'altro, ognuno dei quali ha delle conseguenze se presi a modello per implementare capacità di comprensione sociale nel robot: seguendo la TT, il robot dovrà avere a disposizione un'ampia serie di proposizioni che spiegano le ragioni per cui l'altro compie le proprie azioni, le quali devono essere richiamate all'occorrenza nel processo di dotazione di senso; abbracciando la ST, d'altra parte, il robot dovrebbe essere dotato di capacità di comprensione implicite, basate su meccanismi di risonanza motoria analoghi a quelli realizzati dai neuroni specchio (cfr. §3.3.4).

In questo lavoro di tesi, crediamo di aver fornito gli strumenti concettuali per poter avviare una riformulazione della comprensione di tutti i punti summenzionati, la cui rilettura comporta diverse possibilità per la LfD.

Per quanto riguarda i), nell'ottica enattivista della comprensione dell'altro, la segmentazione dell'azione non ha un ruolo fondamentale. Infatti, il processo al cuore della mutua comprensione è un *attunement* basato su corrispondenze rintracciate al livello corporeo e non inferenziale. Abbiamo visto, infatti, come buona parte delle difficoltà rintracciate da SHAKEY fossero state accantonate sviluppando forme più dinamiche di interazione con l'ambiente e capacità di movimento che non facessero affidamento sulla rappresentazione interna del mondo all'interno del quale il comportamento deve essere realizzato. Da una prospettiva enattivista, infatti, la comprensione non dipende dalla scomposizione dell'azione in parti isolate, ma da una immediata configurazione del pattern di *affordance* e, nel caso della cognizione sociale, dalla connessione diretta e immediata con l'altro. Questo *attunement* si realizza tramite corrispondenze corporee che emergono senza necessità di inferenze rappresentazionali, poiché è il corpo stesso a rintracciare somiglianze attraverso dinamiche percettive e sensorimotorie, rendendo così superfluo un processo di segmentazione analitica.

In merito a ii), nel capitolo §4 abbiamo fornito un'alternativa ai modelli inferenziali della comprensione dell'intenzionalità dell'altro: l'intenzionalità non è inferita a partire dall'osservazione delle azioni dell'altro, ma rintracciata *nel* comportamento, ovvero individuata nell'*engagement* che il sistema intesse con l'ambiente all'interno del quale è inserito e compie le proprie azioni. Così, le azioni dell'altro appaiono da subito significative, in un processo di rilevamento dell'intenzionalità che si basa in prima istanza sulle dinamiche dell'interazione. L'approccio all'imitazione che possiamo

suggerire si basa su una risonanza emotiva ed esperienziale, non sulle inferenze razionalistiche agli stati mentali dell'altro.

Inoltre, alla scelta tra TT e ST al punto iii), abbiamo risposto uscendo dall'*empasse* dato dalla ToM e fornendo una visione dinamica e complessa della cognizione sociale basata sull'interazione (IT) e sulla reciproca coordinazione realizzata sul ritmo dello scambio interindividuale (PSM). Le soluzioni di implementazione basate sulle due versioni più accreditate della ToM, infatti, non sembrano fornire soluzioni agili, né essere completamente disimplicate da una comprensione implicita di tipo corporeo (cfr. Gallagher 2007, 2012). Infatti, la conversione delle idee della ToM in principi alla base dell'architettura cognitiva del robot mostrano importanti difficoltà: la TT comporta un carico computazionale molto elevato, che rallenta enormemente le prestazioni computazionali del robot e inficia notevolmente le capacità di realizzazione del compito in tempo reale; i meccanismi di risonanza e simulazioni postulati nella ST, invece, implicano la necessità che si realizzi innanzitutto una comprensione implicita dell'altro come agente intenzionale, capacità ancora preclusa alle macchine artificiali (Gallagher 2007, 2013). Adottando l'IT e il PSM, abbiamo visto come si possa sorpassare la dicotomia tra TT e ST, fornendo la possibilità di pensare all'interazione sociale come un processo di mutuo riconoscimento che avviene attraverso la coordinazione spontanea che si costituisce nel ritmo dell'interazione: la comprensione dell'altro è un fenomeno emergente dall'interazione stessa, non un processo inferenziale che avviene prima dell'azione.

Ora, a fronte di questo ribaltamento di prospettive, possiamo vedere come sia possibile instradare la ricerca nella LfD riconfigurando gli scopi del processo, la metodologia di indagine e i metodi di verifica attraverso cui valutare la performance del robot.

In primo luogo, dotare il robot delle capacità di apprendere il modo di eseguire un'azione dimostrata non si allinea necessariamente allo scopo di replicare la struttura della mente umana nella macchina. L'elaborazione della variante enattivista del metodo sintetico ci ha permesso di spostare il focus dalla sintetizzazione dei meccanismi cognitivi interni ai risultati che l'incontro del sistema con il suo ambiente realizza, con uno sguardo privilegiato ai processi dinamici di aggiustamento e co-regolazione. Nel dominio dell'interazione sociale con il robot, questo implica che i processi imitativi non debbano essere pensati a partire da una riflessione sulle cause intenzionali del processo, ma sugli effetti che l'azione produce in termini di interazione. La metodologia, dunque, si riorganizza per soddisfare lo scopo di raggiungere l'aggiustamento reciproco in interazione, realizzando modalità di coordinazione senza la necessità di un meccanismo inferenziale a guidarne l'esecuzione. Invece che pensare alla precisa segmentazione del comportamento dell'osservatore, il robot deve puntare a sostenere la struttura interazionale imbastita dall'essere umano, occupando le posizioni vuote allestite per il suo intervento.

A questo scopo, la LfD si presenta come una ottima candidata, dal momento che può consentire al robot di apprendere come muoversi *mentre* l'interazione si svolge. Come riportato da Lee (2017: 8): «many possibilities would be open if we consider robots as our active partners like interactive/active learning based research, and then we can take advantage of capabilities of robot themselves for teaching them». L'apprendimento attivo e interattivo permette di situare la pratica all'interno di uno scambio reciproco e instradare l'apprendimento del robot mentre l'interazione si costituisce. Questo sguardo consente anche di non dover necessariamente rendere il robot consapevole delle condizioni preliminari allo svolgimento dell'azione in interazione, ma di interessare pattern di azioni in risposta a possibilità aperte dall'interazione. In sostanza, dotare il robot della consapevolezza insita nello svolgimento delle nostre pratiche interazionali con l'ambiente e con altri agenti si profila come un progetto a lungo termine basato sulla scommessa che i processi cognitivi alla base della mente umana possano essere modellizzati in strutture computazionali omologhe. Quello che proponiamo qui, invece, è un'indagine a questa complementare, basata su presupposti differenti che, come dimostrato dalla conversione dei principi enattivisti nella robotica situata (§2.2.2), potrebbero portare a vantaggi fino ad ora non considerati.

Un esempio ci può aiutare a chiarire i termini della nostra proposta. Riprendiamo in considerazione la circostanza della stretta di mano che abbiamo più volte analizzato in questo capitolo, e consideriamo in questo momento la struttura di questa interazione. Affinché l'atto sociale della stretta di mano si realizzi, c'è bisogno che entrambi gli agenti compiano una serie di azioni, in un fluido dinamico di gestualità che ha come risultato il movimento della protensione della mano. A questo punto, l'atto viene completato quando un agente si concatena all'atto pragmatico della mano offerta, portando a compimento l'azione iniziata dall'altro. Nelle teorie pragmatiche della comunicazione, azioni come questa vengono definite *atti parziali*: schemi ricorsivi sedimentati a livello sociale di strutture parziali di comportamento, il cui esito necessariamente include un altro agente che completa l'azione (Di Paolo 2021; Di Paolo et al. 2018). La stretta di mano è un esempio paradigmatico, dal momento che l'esito non può realizzarsi se entrambi gli agenti non fanno la loro parte.

Riuscire a rendere capaci i robot di completare questi atti parziali realizzabili solamente nel contesto della socialità costituisce una possibilità a breve termine di creare robot capaci di interagire socialmente con gli esseri umani. I propositi che abbiamo delineato non implicano che il risultato debba essere raggiunto tramite la modellizzazione dei meccanismi cognitivi interni all'agente che, secondo ipotesi, ne consentirebbero la realizzazione, ma che venga raggiunta la struttura ritmica dell'interazione sociale.

I benefici che provengono dalla capacità del robot di concatenarsi negli atti parziali sono già rintracciabili nella letteratura della HRI. Alcuni studi, infatti, testimoniano della capacità

dell'interazione di mitigare effetti spiacevoli come quello dell'UVE realizzando possibilità di concatenazioni dialogiche o agentive (Yam et al. 2021; Paetzel & Castellano 2019), e del maggiore effetto di presenza suscitato da dialoghi in tempo reale (Brainbridge et al. 2008; Grigore et al. 2016).

Con l'avvento delle tecnologie dell'intelligenza artificiale generativa (Sengar et al. 2024), le possibilità di migliorare le performance – sia interazionali, sia dialogiche – senza necessariamente intervenire sulla composizione dell'architettura artificiale interna all'agente robotico si sono fatte ancora più concrete. I primi tentativi di impiego di questo tipo di tecnologie per l'ottimizzazione dell'interazione tra umani e robot sono già in corso (Obrenovic et al. 2024; Fortino & Gaur 2024) e prospettano la possibilità di rendere il robot capace di agire in contesti interazionali e «transform these mechanical devices into interactive, context-aware agents capable of understanding and acting on natural language commands» (Fortino & Gaur 2024: 357). In effetti, i primi video delle prestazioni raggiunte da robot che implementano sistemi di AI generativa come Figure 02, costruito dall'azienda OpenAI<sup>200</sup>, ci permettono di immaginare la possibilità di sviluppare forme molto più complesse di interazione sociale rispetto a quelle che abbiamo visto fino ad ora.

In questa ottica, assume un nuovo potenziale anche il ruolo che la rottura interazionale, in quanto momento di destabilizzazione del piano sensorimotorio e percettivo insito nelle dinamiche di aggiustamento, ha nella pratica di costituzione dell'interazione tra umani e robot. Recentemente, una linea di ricerca negli HRI ha portato alla luce l'importanza dell'errore nell'interazione tra umani e robot proprio come strumento per ottimizzare l'interazione, favorire sviluppi congiunti dell'azione individuale e analizzare gli effetti sulla fiducia dell'essere umano nei confronti del robot e sul potenziale collaborativo percepito (Aroyo et al. 2020; Ye et al. 2019; Kim et al. 2020).

La nostra proposta, dunque, può essere riassunta come il tentativo di esplorare una possibile dimensione della costruzione del senso condiviso, delineata attraverso l'esecuzione di azioni sensorimotorie congiunte al fine di portare a compimento atti parziali. È questa tessitura di azioni a generare il pattern interazionale capace di tenere insieme l'intervento dei singoli, coordinandone ogni frammento. Nella tradizione giapponese, una particolare pratica di riparazione degli oggetti rende bene l'idea della nostra proposta: il *kintsugi*. Utilizzando questa tecnica, infatti, oggetti rotti come tazze e ceramiche vengono recuperati tramite interventi che lasciano ben visibili l'operazione di aggiustamento, addirittura conferendo all'oggetto un valore maggiore dal momento che le riparazioni vengono effettuate con l'oro. Il *kintsugi*, allora, come pratica che cerca di tenere insieme i pezzi, può fornire una metafora programmatica nell'indagine della costruzione delle possibilità interazionali di umani e robot: le rotture non sono solamente momenti di destabilizzazione dell'interazione, ma

---

<sup>200</sup> Un video delle prestazioni di Figure è disponibile su YouTube: <https://www.youtube.com/watch?v=0SRVJaOg9Co>



occasioni da cogliere per aumentare il valore dell'interazione stessa e consentirle di evolvere, trasformarsi, acquisire una propria autonomia attraverso il recupero delle rotture precedenti.

Le prospettive che abbiamo delineato in questa sezione finale, relative alla possibilità di un Metodo Sintetico Enattivista e all'opportunità di inaugurare un'indagine nel campo di studi della LfD, sono solamente due delle conseguenze che possono derivare da una riconfigurazione della HRI sotto il rispetto delle teorie enattiviste della cognizione. Quello che vogliamo sottolineare in chiusura è che includere nell'indagine dell'interazione tra umani e robot il ruolo del corpo, della comprensione situata, della contestualità, della percezione sociale, della coordinazione interindividuale e della reciprocità consente di intravedere nuovi orizzonti e molti potenziali di sviluppo tecnologico. Lungi dal determinare una cesura incolmabile tra scienze della mente e dominio dell'artificiale, questi presupposti possono fornire i principi attraverso cui progettare la costruzione di robot più capaci di comprendere non solo le dinamiche cognitive che permettono l'attraversamento del micromondo individuale, ma soprattutto di intercettare possibilità concrete di concatenamento e coordinazione con l'essere umano. In sintesi, fornire i principi per progettare una *Galatea futura*.

## Conclusioni e rilanci

Alla fine della nostra trattazione, possiamo ora riassumere i risultati che abbiamo raggiunto operando un confronto con i propositi delineati nell'introduzione. Il nostro intento principale è stato quello di verificare se un'indagine filosofica sui presupposti teorici che guidano la concezione del robot, dell'umano e della loro interazione possa condurre a una ridefinizione radicale delle modalità di considerare lo studio della HRI. La nostra ricerca si è basata sulla scommessa che un'operazione di riconfigurazione degli elementi in gioco – dalle teorie sulla mente alle modalità con cui la robotica si è sviluppata congiuntamente alle scienze cognitive, fino all'analisi dei fenomeni centrali per la HRI – potesse aprire nuove prospettive di comprensione del rapporto epistemico che noi umani intratteniamo con i robot e, allo stesso tempo, garantire nuove possibilità di progettazione di architetture cognitive robotiche che possano incorporare i principi provenienti dallo studio particolare dell'interazione con gli esseri umani.

Possiamo provare a fornire, in questo contesto conclusivo, alcune considerazioni riguardo alla creazione di una prospettiva alternativa riguardo lo studio e la comprensione dell'interazione tra umani e robot, che ora possiamo raccogliere a partire dagli esiti del nostro percorso di ricerca.

Dalle riflessioni elaborate nella *Parte I – Storia ed epistemologia della robotica. Sullo sviluppo delle tecniche di mimesi e riproduzione dell'umano*, abbiamo compreso che il robot che incontriamo oggi non è che il risultato di una stratificazione di concetti, sviluppi tecnologici ed evoluzione del processo di conoscenza. Il robot è una macchina particolare, perché le sue capacità di mostrarsi attivo promotore di comportamenti hanno concretizzato – più di tutti gli altri marchingegni e dispositivi tecnologici costruiti nella nostra storia – l'ambizione dell'umano di realizzare una copia di sé stesso, ma questo non deve sfociare in una concezione della macchina come modello totalizzante delle operazioni cognitive. Lo sviluppo della metafora computazionale, reso possibile dalla realizzazione di macchine capaci di sorreggere il peso del ragionamento logico-razionale, ha condizionato l'indagine sulla mente umana facendo passare sottotraccia il fatto che gli assunti che la costituiscono sono il frutto del continuo tentativo di comprendere in qualche modo il mondo che ci circonda. Recuperare la dimensione tentativa e ipotetica delle teorie della mente classiche attraverso il confronto con le alternative filosofiche della mente enattiva e situata consente di mettere in prospettiva l'indagine sulla cognizione, non più vista come un processo di aggiustamento di meccanismi interni che devono necessariamente rispondere alle logiche del computazionalismo. L'alternativa che viene aperta da questa operazione è quella di riconfigurare le possibilità di comprensione situando l'indagine nel contesto pragmatico del rapporto tra l'agente e il suo ambiente. Questa parte ci invita a riflettere sulla possibilità di indagare non solo la struttura interna della cognizione, ma il ruolo che l'interazione tra agenti ha nella costituzione dei processi cognitivi.

L'invito è stato raccolto nella *Parte II – Ti esti. Il robot agli occhi dell'osservatore*, in cui abbiamo indagato le modalità di presentazione che il robot realizza di fronte all'essere umano. La sistematizzazione del campo di ricerca ha rivelato un minimo comun denominatore nella concezione di cognizione sociale sottesa alla totalità degli approcci sviluppati in HRI: la ToM. Aver rintracciato i presupposti epistemologici importati negli studi della HRI dalla teoria della mente ci ha permesso di individuare alcune conseguenze: il connubio tra ToM e HRI ha comportato che l'interazione fosse considerata solamente come il frutto dell'incontro degli agenti umani e robotici, e non un elemento utile nell'indagine dei fenomeni che all'interno si realizzano. La proposta di un approccio alla cognizione sociale alternativo, pensato a partire dall'enattivismo, che mette al centro il ruolo dell'interazione incarnata nell'incontro degli agenti ci ha consentito di sviluppare un'immagine del robot complessa, dinamica e articolata, sotto i diversi aspetti che questo dimostra alla sua controparte dialogica, l'essere umano: agli occhi di chi guarda, il robot si presenta innanzitutto come la simbiosi di due caratteristiche principali, la sincretizzazione realizzata nell'interazione di un aspetto mimetico, che sottolinea la somiglianza estetica, e di uno sintetico, che ne permette l'agentività comportamentale. Questo sguardo ci ha poi permesso di sviluppare un approccio al ruolo dell'intenzionalità in HRI che fosse molto diverso da tutti gli altri presente nel panorama esplicativo attuale: l'*intentionality detection* si è mostrata, allora, come l'effetto armonico dell'interazione con l'umano, che coniuga sotto il rispetto della socialità le caratteristiche esibite dal robot umanoide. Comprendere in maniera dinamica e situata l'emersione delle caratteristiche agentive e interazionali che rendono il robot un potenziale agente intenzionale in interazione ci permette di avanzare nel tentativo di risposta alla domanda iniziale: la progettazione di robot capaci di interagire con l'essere umano deve tenere conto dell'equilibrio degli aspetti robotici nel quadro pragmatico dell'interazione. Nella nostra riflessione, questo significa non pensare alle modalità interazionali in modo astratto, solamente sulla base di teorie circa il funzionamento della cognizione. Ciò che consegue da queste premesse è che l'interazione tra umani e robot che si vuole ottimizzare deve essere considerata nelle sue realizzazioni pragmatiche: solo in questo modo è possibile specificare la necessità di porre l'accento sull'una o sull'altra componente, o tentare di ricreare un'armonia degli effetti agli occhi di chi guarda. Il modo di realizzare robot più efficaci, dunque, passa per le contingenze concrete dell'interazione.

Nella *Parte III – Balla coi robot. L'alterità della macchina*, abbiamo raccolto le osservazioni sviluppate fino a quel momento e le abbiamo rese operative nella costruzione di un approccio enattivista all'interazione tra umani e robot. Lo sviluppo del gioco dell'interazione sulla base della teoria dei micromondi e dei *breakdown* elaborata da Varela ci ha concesso di provare il robot sociale su un terreno di confronto a noi familiare: quello dell'incontro incarnato, concreto e pratico delle

circostanze sociali. Rileggere l'UVE sotto il rispetto del gioco dell'interazione ci ha permesso i) di comprendere come le ipotesi esplicative fino ad ora considerate risentano della stessa influenza della ToM che pervade tutti gli studi di HRI e che determina la considerazione degli effetti in termini razionalistici e componenziali e ii) di strutturare il gioco prendendo in considerazione le teorie che definiscono il procedere dell'interazione in ottica pragmatica e situata. Così, le teorie del *predictive engagement* e del PSM ci hanno fornito gli strumenti epistemologici per analizzare l'UVE nella sua struttura processuale: abiti e aspettative incarnate allestiscono le salienze nel micromondo, mentre la coordinazione reciproca in relazione al ritmo dell'interazione diagrammatizza le possibilità di intervento dell'agente e apre posizioni che potrebbero essere occupate nella costruzione di senso condiviso. I *breakdown* come l'UVE non costituiscono motivo di rotture irreparabili, ma possibilità di accrescimento della conoscenza in merito a ciò che accade effettivamente nell'incontro tra robot e umani. Nelle conseguenze che abbiamo delineato alla fine di questo capitolo possiamo ritrovare un ulteriore sviluppo della risposta alla domanda iniziale: tramite la struttura del gioco dell'interazione, abbiamo capito che il robot non può essere progettato con la speranza che esibisca le nostre stesse capacità interazionali. La nostra abilità di surfare sulle onde dell'interazione ci è data dalla costitutiva precarietà della nostra cognizione e dalle capacità che questa precarietà concede, sia a livello percettivo che agentivo. Pensare di poter replicare la capacità dell'umano di adattarsi a contesti così cangianti, pieni di rotture che necessitano recuperi, che mostrano una costitutiva instabilità del piano di intervento sembra essere un obiettivo, allo stato attuale, troppo ambizioso per la robotica sociale. Tuttavia, le riflessioni che abbiamo compiuto aprono nuove strade per la progettazione di robot interazionali: la proposta di un metodo sintetico basato sugli approcci enattivisti e della sua possibile inclusione nelle tecniche di *Learning from Demonstration* si costituisce proprio a partire dalla possibilità di modellizzare interazioni specifiche, individuando le contingenze e le possibilità di coordinazione, su tanti piani quanti sono quelli che il robot può navigare.

Riteniamo, allora, che solo attraverso uno sguardo filosofico diverso rispetto a quello classico, che rivoluzioni strutturalmente il modo che abbiamo di pensare alla cognizione umana, sia possibile impostare la riflessione sulla robotica che verrà, adottando i modelli dell'interazione che abbiamo cercato di costruire: quelli della collaborazione, del dialogo e dell'alterità. La cognizione robotica e quella umana non funzionano secondo gli stessi principi – crediamo di averlo mostrato – ma questo non può portare a un abbandono dell'impresa e alla dichiarazione di una radicale in traducibilità tra i due sistemi. Al contrario, questi sistemi possono essere studiati in interazione per migliorare la comprensione e le prestazioni di entrambi gli agenti coinvolti. Da una parte questa ricerca, delineando e rintracciando attraverso una metodologia interdisciplinare fondata su studi filosofici e cognitivi gli elementi fondamentali della capacità interazionale dell'essere umano, potrebbe permettere la

realizzazione e l'implementazione di architetture robotiche che potranno sfruttare queste stesse caratteristiche per una maggiore efficacia interazionale; d'altra parte e allo stesso tempo, proprio questo sforzo potrà portare a una maggiore comprensione delle modalità cognitive umane, rafforzando o indebolendo le ipotesi teoriche – nate nell'alveo della filosofia – attraverso un'indagine empirica sull'interazione tra gli umani e i robot costruiti sulla base delle stesse. Se dovessimo scommettere su uno dei possibili esiti di tale implementazione, esso sarebbe la realizzazione di robot più funzionali a determinati scopi, non necessariamente migliori di altri, né più simili all'umano o all'altezza di sostituirlo, ma certamente più capaci di svolgere con noi compiti che sinora si dimostrano irrealizzabili.

In conclusione, e proprio in riferimento a questo aspetto, vogliamo esplicitare un pensiero che riassume l'intento di questo lavoro. La *Galatea futura* che abbiamo immaginato non si profila come più capace di noi nelle interazioni sociali, né più efficiente nei processi di costruzione di senso rispetto a quanto già siamo. Essa non è concepita come una replica idealizzata dell'umano, ma come una interlocutrice che, dotata di una voce propria, possa instaurare con noi un dialogo prolifico. In questa prospettiva, il robot non è *altro da noi* con l'intento di sostituirci, ma una alterità presente che ci permette di riarticolare le nostre domande, di riflettere sulle modalità della nostra interazione con l'altro e, in fondo, con noi stessi. Così intesa, questa *Galatea* si rivela uno specchio capace di restituirci non solo un'immagine di noi stessi, ma un riflesso che ci interroga e ci stimola a riconsiderare il nostro approccio all'alterità. I robot, lungi dall'essere destinatari finali delle nostre aspettative, possono diventare compagni di esplorazione in un processo reciproco di co-costruzione del senso, aiutandoci, insieme, a immaginare nuove *Galatee future*. Attraverso questa prospettiva, si apre un sentiero verso una robotica che non abbia come obiettivo quello di costruire macchine che imitino *semplicemente* l'umano, ma che possano contribuire a un'inedita comprensione dell'interazione, dell'intelligenza e del nostro rapporto con gli artefatti che realizziamo, rendendo il robot un partner interazionale attivo e, al tempo stesso, una possibilità aperta su ciò che, come esseri umani, ancora desideriamo conoscere. In questo panorama, tuttavia, nessuna possibilità di comprensione è data per scontata e, anzi, forse dice bene il personaggio di Edison nato dalla penna di Villiers de l'Isle-Adam, afferrare la chiave di fenomeni insoliti e impressionanti non costituisce garanzia di intendimento. Crediamo fermamente che sia in questo spazio di incertezza che si muove l'opera di conoscenza, lasciando aperti spiragli di intelligibilità anche quando, all'apparenza, possa sembrare di aver indagato a fondo il nostro oggetto di ricerca.

Dunque, con una dichiarazione di incapacità nello spiegare la natura di ciò che fa muovere i meccanismi della vita cominciava il discorso sui robot del saggio Edison; con la stessa dichiarazione

si chiude il nostro: che avremo meno saggezza, ma la stessa fascinazione per ciò che ancora è di là da conoscere.

## Bibliografia

Nel testo, l'anno che accompagna i rinvii bibliografici è quello dell'edizione in lingua originale, mentre i rimandi ai numeri di pagina si riferiscono alla traduzione italiana o inglese, qualora sia presente nella bibliografia.

- Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in psychology*, 8, 1393.
- Abubshait, A., & Wykowska, A. (2020). Repetitive robot behavior impacts perception of intentionality and gaze-related attentional orienting. *Frontiers in Robotics and AI*, 7, 565825.
- Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14(1), 43–64.
- Adamson, L. B., & Frick, J. E. (2003). The still face: A history of a shared experimental paradigm. *Infancy*, 4(4), 451-473.
- Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1), 25-63.
- Agre, P. E. (1985). Robotica e buon senso. In Minsky, M. (Ed.) (1985) *Robotics: The First Authoritative Report from the Ultimate High-Tech Frontier*. New York Doubleday. (tr. it. Sosio, L., *La robotica. Il primo autorevole rapporto delle frontiere dell'alta tecnologia*. Milano: Longanesi, 1987, 71-100).
- Agre, P. E. (1988). *The dynamic structure of everyday life*. Cambridge, MA: MIT Artificial Intelligence Laboratory.
- Agre, P. E., & Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the sixth National conference on Artificial intelligence - Volume 1*, 268-272.
- Aguilar, W., Santamaría-Bonfil, G., Froese, T., & Gershenson, C. (2014). The past, present, and future of artificial life. *Frontiers in Robotics and AI*, 1, 8.
- Airenti, G. (2015). The cognitive bases of anthropomorphism: from relatedness to empathy. *International Journal of Social Robotics*, 7, 117-127.
- Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in psychology*, 9, 2136.
- Akdeniz, G., Deliktaş, Ö. D., Fincan, A. C., Yilmaz, G., Yoldaş, T., & ÖZİŞİK, P. (2024). Examining the relationship between touch and visual perception: pareidolia perception in the social brain. *Pamukkale Medical Journal*, 17(2), 3-3.
- Akechi, H., Senju, A., Uiho, H., Kikuchi, Y., Hasegawa, T., & Hietanen, J. K. (2013). Attention to eye contact in the West and East: Autonomic responses and evaluative ratings. *PloS one*, 8(3), e59312.
- Akhtaruzzaman, M., & Shafie, A. A. (2010). Evolution of humanoid robot and contribution of various countries in advancing the research and development of the platform. In *ICCAS 2010*, 1021-1028, IEEE.

- Allen, M. & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, 195, 2459–2482, doi:10.1007/s11229-016-1288-5.
- Allen, M., & Tsakiris, M. (2018). The body as first prior: Interoceptive predictive processing and the primacy. *The interoceptive mind: From homeostasis to awareness*, 27.
- Alliney, G. (Ed.) (1951) *Gli occultisti*. Milano: Garzanti.
- Alvarez Perez, J., Garcia Goo, H., Sánchez Ramos, A., Contreras, V., & Strait, M. (2020, March). The uncanny valley manifests even with exposure to robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 101-103).
- Amirova, A., Rakhymbayeva, N., Yadollahi, E., Sandygulova, A. and Johal, W. (2021). 10 Years of Human-NAO Interaction Research: A Scoping Review. *Front. Robot. AI* 8:744526. doi: 10.3389/frobt.2021.744526
- Andler, D. (2023). *Intelligence artificielle, intelligence humaine: la double énigme*. Paris: Gallimard (tr. it. *Il duplice enigma: Intelligenza artificiale e intelligenza umana*, Torino, Einaudi, 2024).
- Andreasson, R., Alenljung, B., Billing, E., & Lowe, R. (2018). Affective touch in human–robot interaction: conveying emotion to the Nao robot. *International Journal of Social Robotics*, 10, 473-491.
- Andrew, A. M. (2005). Artificial neural nets and BCL. *Kybernetes*, 34(1/2), 33-39.
- Andrews-Hanna, J. R. (2012). The brain’s default network and its adaptive role in internal mentation. *The Neuroscientist*, 18(3), 251-270.
- Angelucci, A., Rossi, M. G., & Graziani, P. (2016). The Uncanny Valley. A Working Hypothesis. In *Social Robots: Boudaries, Potential, Challenges* (pp. 123-137). Routledge.
- Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.
- Ansuini, C., Cavallo, A., Bertone, C., & Becchio, C. (2014). The visible face of intention: why kinematics matters. *Frontiers in psychology*, 5, 815.
- Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior*, 102, 274–286. <https://doi.org/10.1016/j.chb.2019.08.018>
- Apperly, I. A. (2011). *Mindreaders: The cognitive basis of “Theory of Mind”*. Hove: Psychology Press.
- Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, 65(5), 825–839.
- Arango, A. (2019). From sensorimotor dependencies to perceptual practices: Making enactivism social. *Adaptive behavior*, 27(1), 31-45.
- Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5), 469-483



- Arkin, R. C. (1998). *Behavior-based robotics*. Cambridge (MA): MIT press.
- Arora, A. S., Sivakumar, K., Taras, V. (2024) The role of anthropomorphic, xenocentric, intentional, and social (AX<sup>IS</sup>) robotics in human-robot interaction. *Computers in Human Behavior: Artificial Humans*, Volume 2, Issue 1, 2024.
- Aroyo, A. M., Pasquali, D., Kothig, A., Rea, F., Sandini, G., & Sciutti, A. (2020). Perceived differences between on-line and real robotic failures. In *Proc. SCRITA Workshop-Trust, Acceptance Social Cues Hum.-Robot Interaction*.
- Ashby W. R. (1947) Principles of the Self-Organizing Dynamic System. *The Journal of General Psychology*, 37:2, 125-128.
- Ashby, W. R. (1946). Dynamics of the cerebral cortex: the behavioural properties of systems in equilibrium. *The American journal of psychology*, 59(4), 682-686.
- Ashby, W. R. (1952). *Design for a Brain. The Origin of Adaptive Behavior*. New York: John Wiley and Sons Inc. (trad. it. *Progetto per un cervello. L'origine del comportamento adattivo*, Orthotes, Napoli-Salerno, 2021).
- Asimov, I. (1940) Robbie. *Super Science Stories* (tr. it. In *Io, Robot*, Mondadori, Milano, 2021, pp. 13-36)
- Asimov, I. (1950) *I, Robot*. New York: Doubleday (tr. it. *Io, Robot*, Milano, Mondadori, 2018).
- Asimov, I. (1990) Introduction: The Robot Chronicles. In *Robot Visions*. Byron Preiss Visual Publications (tr. it. *Visioni di robot*, il Saggiatore, Milano, 2019)
- Asimov, I., & Frenkel, A. (1985). *Robots: Machines in Man's Image*. New York: Harmony Books.
- Augey, D. & Alcaraz, M. (2019). Are Robots and AI the Future of the Media?. In *Digital Information Ecosystems: Smart Press*, Wiley, 2019, pp.161-177, doi: 10.1002/9781119579717.ch8.
- Avramides, A. (2023). Other Minds. In *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2023/entries/other-minds/>
- Bacaro, M. (forthcoming). The Robot as an Icon: Exploring Ambiguity in Perception. In un volume a cura di Dondero, M. G., Pinotti, A., Dalmasso, A. C. & Conte, P. (Eds.) *AlgoMedia. The Image at the Time of Artificial Intelligence*. Springer, serie «Lecture Notes in Morphogenesis ».
- Bacaro, M. (in press). Why Do Robots Have Faces? A Cognitive Semiotic Perspective on Human-Likeness in Robots, in Leone, M., Dondero, M.G., Paolucci, C. *Face it! The new challenges of Cognitive Visual Semiotics*, Springer, Dordrecht.
- Bacaro, M. & Bianchini, F. (2024). Artificial Intelligence ad Expected Intelligence, in Bianchini, F., Fano, V., Graziani, P. (Eds.) *Current Topics in Logic and the Philosophy of Science. Papers from SILFS 2022 postgraduate conference*. College Publications, pp. 89-116.
- Bacone, F. *De secretis operibus artis et naturae*, IV (tr. it. *I segreti dell'arte e della natura*, in Bacone, F. (a cura di Bottin, F.) *La scienza sperimentale*, Milano, Rusconi, 1990)

- Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3, 41-52.
- Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 701-706). IEEE.
- Bakan, U. (2021). Examining the uncanny valley effect in virtual character design for digital games. *Journal of Print and Media Technology Research*, 10(2), 119-132.
- Balkenius, C., & Johansson, B. (2022). Almost alive: Robots and androids. *Frontiers in Human Dynamics*, 4, 703879.
- Banks, J. (2020). Theory of mind in social robots: replication of five established human tests. *International Journal of Social Robotics*, 12(2), 403-414.
- Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive behavior*, 17(5), 367-386.
- Barchard, K. A., Lapping-Carr, L., Westfall, R. S., Fink-Armold, A., Banisetty, S. B., & Feil-Seifer, D. (2020). Measuring the perceived social intelligence of robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(4), 1-29.
- Bardi, L., Regolin, L., & Simion, F. (2011). Biological motion preference in humans at birth: Role of dynamic and configural properties. *Developmental science*, 14(2), 353-359.
- Baron-Cohen S., Leslie A. M., & Frith U. (1985). Does the autistic child have a “theory of mind”? *Cognition* 21 37–46. 10.1016/0010-0277(85)90022-8
- Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, 1(233-251), 1.
- Baron-Cohen, S. (1993). The concept of intentionality: Invented or innate?. *Behavioral and Brain Sciences*, 16(1), 29-30.
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT press.
- Baron-Cohen, S. (1999). Evolution of a theory of mind?. In Corballis, M., & Lea, S (eds), *The descent of mind: psychological perspectives on hominid evolution*. Cambridge (MA): Oxford University Press
- Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International review of research in mental retardation*, 23, 169-184.
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2), 241-251.

- Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1325-1334.
- Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature reviews neuroscience*, 16(7), 419-429.
- Bartneck, C. (2003). Interacting with an embodied emotional character. In *Proceedings of the 2003 international conference on Designing pleasurable products and interfaces*(pp. 55-60).
- Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759)* (pp. 591-594). IEEE.
- Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, 9(3), 415-433.
- Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., & Šabanović, S. (2020). *Human-robot interaction: An introduction*. Cambridge, MA: Cambridge University Press.
- Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2007). Is the uncanny valley an uncanny cliff?. In *RO-MAN 2007-The 16th IEEE international symposium on robot and human interactive communication* (pp. 368-373). IEEE.
- Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009,). My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication* (pp. 269-276). IEEE.
- Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1, 71-81.
- Bartneck, C., Verbunt, M., Mubin, O., & Al Mahmud, A. (2007, March). To Kill a Mockingbird Robot. In *Proceedings of the ACM/IEEE international conference on human-robot interaction*, (pp. 81-87).
- Bastian, B., & Haslam, N. (2011). Experiencing dehumanization: Cognitive and emotional effects of everyday dehumanization. *Basic and Applied Social Psychology*, 33(4), 295-303.
- Battistini, A. (2015) Il mito di Prometeo in età moderna: dal peccato di «hybris» alla virtù della «curiositas». In *Le «borie» vichiane come paradigma euristico*. Atti del convegno di Napoli (22-23 novembre 2012) 191-208.
- Baumgartner, P., & Payr, S. (Eds.). (1996). *Speaking minds: Interviews with twenty eminent cognitive scientists*. Princeton: Princeton University Press.
- Baxter, P. E., de Greeff, J., & Belpaeme, T. (2013) Cognitive architecture for human–robot interaction: Towards behavioural alignment. *Biologically Inspired Cognitive Architectures*, Volume 6, 2013, Pages 30-39, <https://doi.org/10.1016/j.bica.2013.07.002>.
- Becchio, C., Sartori, L., Bulgheroni, M., & Castiello, U. (2008). Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition*, 106(2), 894-912.

- Bechtel, W. (2003). Modularity in Cognitive Science: Seeking Underpinnings for a New Research Program. In W. Bechtel & R. Graham (Eds.), *A Companion to Cognitive Science* (pp. 635–644). Blackwell.
- Bedaf, S., Huijnen, C., van den Heuvel, R., & de Witte, L. (2017). Robots supporting care for elderly people. In *Robotic assistive technologies* (pp. 309-332). CRC Press.
- Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of human-robot interaction*, 3(2), 74.
- Beer, R. D. (1990). *Intelligence As Adaptive Behavior. An Experiment in Computational Neuroethology*. San Diego (CA): Academic Press.
- Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, 20(2-4), 257-289.
- Beer, R. D. (2023). On the proper treatment of dynamics in cognitive science. *Topics in cognitive science*. <https://doi.org/10.1111/tops.12686>
- Bello, A. A. (2005). Husserl interprete di Kant. *Dialegesthai*, 7. <https://purl.org/mdd/angela-ales-bello-02>
- Ben-Ari, M., Mondada, F. (2018). Finite State Machines. In *Elements of Robotics*. Springer, Cham.
- Benjamin, R., & Heine, S. J. (2023). From freud to android: Constructing a scale of uncanny feelings. *Journal of personality assessment*, 105(1), 121-133.
- Bennett, J. (1978). Some remarks about concepts. *The Behavioral and Brain Sciences* 1: 557–560.
- Benninghoff, B., Kulms, P., Hoffmann, L., & Krämer, N. C. (2013). Theory of mind in human-robot-communication: Appreciated or not?. *Kognitive Systeme*, 2013(1).
- Bering, J. (2005) *Origins of social mind*. New York, NY: The Guilford Press.
- Bermudez, J. L. (2020) *Cognitive Science. An Introduction to the Science of Mind* (3rd edition). Cambridge, MA: Cambridge University Press.
- Bernardi, W. (2014) Meccanicismo e Vitalismo. In Eco, U., Corsi, P. (Eds.) *Il Settecento - Scienze e tecniche*. Milano: EM Publishers.
- Bernotat, J., & Eyssel, F. (2018). Can ('t) wait to have a robot at home?-Japanese and German users' attitudes toward service robots in smart homes. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 15-22). IEEE.
- Bernstein, J. (1978/1982) *Experiencing Science/Science Observed*. New York: Basic Books (tr. it. *Uomini e machine intelligenti*, Milano, Adelphi, 1990)
- Berry, D. S., & Springer, K. (1993). Structure, motion, and preschoolers' perceptions of social causality. *Ecological Psychology*, 5(4), 273-283.
- Bertalanffy, L., von (1951). Towards a physical theory of organic teleology. *Human biology*, 23(4), 346.

- Bertalanffy, L., von (1964). The mind-body problem: A new view. *Psychosomatic Medicine*, 26(1), 29-45.
- Bertalanffy, L., von (1968) *General System Theory. Foundations, development, applications*. New York: Braziller.
- Bianchini, F. (2014) Turing and the evaluation of intelligence. *2014 Isonomia, Rivista online di Filosofia – Epistemologica*, 1-18.
- Bianchini, F. (2023). Autopoiesis of the artificial: from systems to cognition. *Biosystems*, 230, 104936.
- Bianchini, F., Damiano, L., Datteri, E. et al. (2023) Experimental and Integrative Approaches to Robo-ethics. An Introduction. *Int J of Soc Robotics* 15, 1977–1980 (2023). <https://doi.org/10.1007/s12369-023-01084-z>
- Bianchini, M., Simic, M., Ghosh, A., & Shaw, R. N. (2021). Machine Learning for Robotics Applications. *Studies in Computational Intelligence*, 960.
- Bianco, F., & Ognibene, D. (2019a). Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings* 11 (pp. 77-87). Springer International Publishing.
- Bianco, F., & Ognibene, D. (2019b). Bianco, F., & Ognibene, D. (2019). Functional advantages of an adaptive theory of mind for robotics: a review of current architectures. *2019 11th Computer Science and Electronic Engineering (CEECE)*, 139-143.
- Bisconti, P., & Carnevale, A. (2022). Alienation and Recognition-The  $\Delta$  Phenomenology of Human-Social Robot Interactions. *Techné: Research in Philosophy and Technology*, 26(1).
- Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PloS one*, 9(8), e106172.
- Bizzari, V., & Hipolito, I. (2016). Predictive engagement and motor intentionality. *Esercizi Filosofici* 11, 2016, pp. 86-98
- Block, N. (1987). Advertisement for a Semantics for Psychology. *Midwest studies in philosophy*, 10(1), 615-678.
- Block, N., (1995). The mind as the software of the brain. In Smith, E.E., Osherson, D.N. (Eds.), *Thinking: An Invitation to Cognitive Science*. Cambridge, MA: MIT Press, 377–425.
- Blow, M., Dautenhahn, K., Appleby, A., Nehaniv, C. L., & Lee, D. C. (2006). Perception of robot smiles and dimensions for human-robot interaction design. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 469-474). IEEE.
- Bocchi, G. & Ceruti, M. (Eds) (1985). *La sfida della complessità*. Milano: Feltrinelli.
- Boden, M. (2014). GOFAI. In Frankish, K. e Ramsey, W. M. (Eds) *The Cambridge Handbook of Artificial Intelligence*, Cambridge, MA: Cambridge University Press, 89-107.

- Boden, M. (1977) *Artificial Intelligence and Natural Man*. New York: Basic Books).
- Boden, M. (2006). *Mind as Machine. A history of Cognitive Science*. Cambridge, MA: Cambridge University Press.
- Bongard, J. C. (2013). Evolutionary robotics. *Communications of the ACM*, 56(8), 74-83.
- Borges, J. L. (1937), La macchina per pensare di Raimondo Lullo. In *Testi prigionieri*, Milano, Adelphi, 1998.
- Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V., & Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science Robotics*, 5(46), Article eabb6652. <https://doi.org/10.1126/scirobotics.abb6652>
- Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford (MA): Oxford University Press.
- Boyer, P. (1996). What makes anthropomorphism natural: Intuitive ontology and cultural representations. *Journal of the Royal Anthropological Institute*, 83-97.
- Brady, J.M. (1985) Artificial Intelligence and Robotics. *Artificial Intelligence*, 26: 79-122.
- Braitenberg, V. (1984). *I veicoli pensanti. Saggio di psicologia sintetica*. Milano: Garzanti.
- Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.
- Brandom, R. (2000) *Articulating Reasons: An introduction to inferentialism*. Cambridge (MA): Oxford University Press.
- Breazeal, C. (2002). Regulation and entrainment in human—robot interaction. *The International Journal of Robotics Research*, 21(10-11), 883-902.
- Breazeal, C. (2003a). Emotion and Sociable Humanoid Robots. *International Journal of Human-Computer Studies*, 59(1-2), 119-155.
- Breazeal, C. (2003b). Toward sociable robots. *Robotics and Autonomous Systems*, 42 (3-4), 167-175.
- Breazeal, C. (2004a). *Designing sociable robots*. Cambridge, MA: MIT press.
- Breazeal, C. (2004b). Social interactions in HRI: the robot view. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, 34(2), 181-186.
- Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in cognitive sciences*, 6(11), 481-487.
- Breazeal, C., Dautenhahn, K., Kanda, T. (2016). Social Robotics. In Siciliano, B., Khatib, O. (eds) *Springer Handbook of Robotics*. Springer Handbooks. Springer, Cham.
- Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Leipzig. (tr. it. *La psicologia dal punto di vista empirico*, a cura di C. Sini, Laterza, Roma-Bari, 1997).
- Brenton, H., Gillies, M., Ballin, D., & Chatting, D. (2005). The Uncanny Valley: does it exist and is it related to presence. *Presence connect*, 8(1).



- Brenton, H., Gillies, M., Surman, D., & Sloan, H. (2007). Uncanny Valley: Realism In Visualisation Character Design. *SwanQuake: The User Manual*, 93-100.
- Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child development*, 90(4), 1202-1214.
- Broadbent, E. (2017). "Interactions with robots: The truths we reveal about ourselves." *Annual Review of Psychology*, 68, 627-652.
- Broadbent, E., Kumar, V., Li, X., Sollers 3rd, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS one*, 8(8), e72589.
- Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2), 94-103.
- Brooks, R. (1990) Elephants don't play chess. *Robotics and Autonomous Systems* 6, 3-15.
- Brooks, R. (1986) Achieving Artificial Intelligence Through Building Robots. *Technical Report Artificial Intelligence Laboratory, Massachusetts Institute of Technology*: Cambridge, MA.
- Brooks, R. (1989). The whole iguana. *Robotics science*, 432-456.
- Brooks, R. (1991a) Intelligence without reason. *Technical Report* [Computers and Thought, IJCAI-91]; *Artificial Intelligence Laboratory, Massachusetts Institute of Technology*: Cambridge, MA.
- Brooks, R. (1991b). New approaches to robotics. *Science*, 253(5025), 1227-1232.
- Brooks, R. (1991c) Intelligence without representation. *Artificial Intelligence* 47, 139-159.
- Brooks, R. (1997). From earwigs to humans. *Robotics and autonomous systems*, 20(2-4), 291-304.
- Brooks, R. (2002) *Flesh and Machines. How Robots Will Change Us*. New York: Vintage Books.
- Brooks, R. A., & Stein, L. A. (1994). Building brains for bodies. *Autonomous Robots*, 1, 7-25.
- Brooks, R. A., Breazeal, C., Marjanović, M., Scassellati, B., & Williamson, M. M. (1998). The Cog project: Building a humanoid robot. In *International workshop on computation for metaphors, analogy, and agents* (pp. 52-87). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Bruin, de L. D., & Haan, S. D. (2012). Enactivism & social cognition: In search of the whole story. *Cognitive Semiotics*, 4(1), 225-250.
- Bruineberg, J., Kiverstein, J. and Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417–2444. doi:10.1007/s11229-016-1239-1.
- Bruinenberg, J. (2018). *Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems*. PhD Thesis, University of Amsterdam, Institute for Logic, Language and Computation (ILLC). <https://hdl.handle.net/11245.1/75672790-f4ac-4b84-bde7-83b51ccc73bb>
- Brüne, M., Abdel-Hamid, M., Lehmkämpfer, C., & Sonntag, C. (2007). Mental state attribution, neurocognitive functioning, and psychopathology: what predicts poor social competence in schizophrenia best?. *Schizophrenia research*, 92(1-3), 151-159.

- Bruneau, E., & Kteily, N. (2017). The enemy as animal: Symmetric dehumanization during asymmetric warfare. *PloS one*, 12(7), e0181422.
- Bryson, J. J. (2010). Robots should be slaves. *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, 8, 63-74.
- Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, 4, 1094.
- Buhrmann, T., & Di Paolo, E. (2017). The sense of agency—a phenomenological consequence of enacting sensorimotor schemes. *Phenomenology and the Cognitive Sciences*, 16, 207-236.
- Burgoon, J. K., & Bacue, A. E. (2003). Nonverbal communication skills. *Handbook of communication and social interaction skills*, 179-219.
- Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in human behavior*, 29(3), 759-771.
- Byom LJ, Mutlu B. (2013). Theory of mind: mechanisms, methods, and new directions. *Front Hum Neurosci*. 2013 Aug 8; 7: 413. doi: 10.3389/fnhum.2013.00413. PMID: 23964218; PMCID: PMC3737477.
- Calinon, S. (2018). Learning from Demonstration (Programming by Demonstration). In: Ang, M., Khatib, O., Siciliano, B. (eds) *Encyclopedia of Robotics*. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-41610-1\\_27-1](https://doi.org/10.1007/978-3-642-41610-1_27-1)
- Cangelosi, A. (2017). Connectionist and robotics approaches to grounding symbols in perceptual and sensorimotor categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 795–818). Elsevier Academic Press.
- Cangelosi, A., & Asada, M. (Eds.) (2022). *Cognitive robotics*. Cambridge, MA: MIT Press.
- Cangelosi, A., & Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press.
- Capek, K. (1921). *Rossum's universal robots (RUR)*. Penguin, London Castells M (2000) The rise of the network society, 2nd edn. Blackwell, Malden Cooley M (2007) From judgment to calculation. *AI Soc*, 21(4), 395-409 Cowley.
- Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in human behavior*, 2(3), 215-234.
- Cappuccio, M. (2005). *Alan Turing: l'uomo, la macchina, l'enigma; per una genealogia dell'incomputabile*. Milano: Alboversorio.
- Capra, F. & Luisi, P. (2016) *The system view of life*. Cambridge, MA: Cambridge University Press (tr. it. *Vita e natura. Una visione sistemica*. Arezzo: Aboca.
- Carlucci Aiello, L., Nardi, D., & Pirri, F. (2012). Case studies in cognitive robotics. In *Human and Machine Perception 3: Thinking, Deciding, and Acting* (pp. 167-181). Boston, MA: Springer US.



- Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*(pp. 254-262).
- Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.
- Carter, E. J., Hodgins, J. K., & Rakison, D. H. (2011). Exploring the neural correlates of goal-directed action and intention understanding. *NeuroImage*, 54(2), 1634–1642.
- Caruana, N., & Seymour, K. (2022). Objects that induce face pareidolia are prioritized by the visual system. *British Journal of Psychology*, 113(2), 496-507.
- Casari, E. (1973). *La filosofia della matematica del '900*. Firenze: Sansoni.
- Castelo, N., Sarvary, M. (2022). Cross-Cultural Differences in Comfort with Humanlike Robots. *Int J of Soc Robotics* 14, 1865–1873 (2022). <https://doi.org/10.1007/s12369-022-00920-y>
- Ceruti M. & Damiano, L. (2013). Embodiment enattivo, cognizione e relazione dialogica. *Encyclopaideia* XVII (37), 19-46.
- Ceruti, M. (1985) La hybris dell'onniscienza e la sfida della complessità. In Bocchi, G. e Ceruti, M. (Eds) (1985) *La sfida della complessità*. Milano: Feltrinelli, 25-48.
- Ceruti, M. (1989). *La danza che crea. Evoluzione e cognizione nell'epistemologia genetica*. Milano: Feltrinelli.
- Ceruti, M., & Damiano, L. (2018). Plural embodiment (s) of mind. genealogy and guidelines for a radically embodied approach to mind and consciousness. *Frontiers in Psychology*, 9, 413623.
- Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.
- Chaminade, T, D. W. Franklin, E. Oztop and G. Cheng, (2005) "Motor interference between Humans and Humanoid Robots: Effect of Biological and Artificial Motion," *Proceedings. The 4th International Conference on Development and Learning, 2005*, Osaka, Japan, 2005, pp. 96-101, doi: 10.1109/DEVLRN.2005.1490951.
- Chaminade, T., & Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *Journal of Physiology-Paris*, 103(3-5), 286-295.
- Chaminade, T., Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., & Prévot, L. (2018). Investigating the dimensions of conversational agents' social competence using objective neurophysiological measurements. In *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct* (pp. 1–7).
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, 6, 103.
- Chan et al. (2012). Learning in Robots. In: Seel, N.M. (eds) *Encyclopedia of the Sciences of Learning*. Springer, Boston, MA. [https://doi.org/10.1007/978-1-4419-1428-6\\_4674](https://doi.org/10.1007/978-1-4419-1428-6_4674)

- Chanes, L., & Barrett, L. F. (2020). The predictive brain, conscious experience, and brain-related conditions. *The philosophy and science of predictive processing, 1*, 159-169.
- Changeux, P. (1983) *L'homme neuronal*. Paris: Fayard.
- Cheetham, M. (Ed.). (2018). *The uncanny valley hypothesis and beyond*. Frontiers Media SA.
- Cheetham, M., Suter, P., & Jancke, L. (2014). Perceptual discrimination difficulty and familiarity in the uncanny valley: more like a “Happy Valley”. *Frontiers in psychology, 5*, 1219.
- Cheetham, M., Wu, L., Pauli, P., & Jancke, L. (2015). Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. *Frontiers in psychology, 6*, 981.
- Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
- Chien, S. E., Chen, Y. S., Chen, Y. C., & Yeh, S. L. (2024). Exploring the Developmental Aspects of the Uncanny Valley Effect on Children’s Preferences for Robot Appearance. *International Journal of Human–Computer Interaction*, 1-11.
- Chin, M. G., Sims, V. K., Clark, B., & Lopez, G. R. (2004). Measuring individual differences in anthropomorphism toward machines and animals. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 48, No. 11, pp. 1252-1255). Sage CA: Los Angeles, CA: SAGE Publications.
- Chisholm, R. M. (1973). Beyond being and nonbeing. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 24*(4), 245-257.
- Chua, H. F., Boland J. E., Nisbett R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 12629–12633.
- Ciardo, F., De Tommaso, D., & Wykowska, A. (2022). Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test. *Science robotics, 7*(68), eabo1241.
- Ciaunica, A., Constant, A., Preissl, H., & Fotopoulou, K. (2021a). The first prior: from co-embodiment to co-homeostasis in early life. *Consciousness and cognition, 91*, 103117.
- Ciaunica, A., Safron, A., Delafield-Butt, J. (2021b). Back to square one: the bodily roots of conscious experiences in early life. *Neuroscience of Consciousness*, Volume 2021, Issue 2, 2021, niab037, <https://doi.org/10.1093/nc/niab037>
- Clapin, H. (2002) *Philosophy of Mental Representation*. Oxford: Oxford University Press.
- Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT press.
- Clark, A. (1999). An embodied cognitive science?. *Trends in cognitive sciences, 3*(9), 345-351. doi: 10.1016/s1364-6613(99)01361-3
- Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. oxford university Press.

- Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, 121(483), 753-771.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), 181-204.
- Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, NY
- Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7-19. <http://www.jstor.org/stable/3328150>
- Coeckelbergh, M. (2010) Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12(3):209-221.
- Coeckelbergh, M. (2011) Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *Int J of Soc Robotics* 3, 197-204. <https://doi.org/10.1007/s12369-010-0075-6>
- Cohen, J. (1966) *Human Robots in Myth and Science*. George Allen & Unwin Ltd. (tr. it. *I ROBOT nel mito e nella scienza*. Bari: De Donato, 1980)
- Collins, H. M. (2012). Expert systems and the science of knowledge. In *The social construction of technological systems: new directions in the sociology and history of technology*. Boston: MIT Press, 321-348.
- Colombetti, G. (2013). Psychopathology and the Enactive Mind. In K. W. M. Fulford, and others (eds), *The Oxford Handbook of Philosophy and Psychiatry*, pp. 1083-1102 (2013; online edn, Oxford Academic, 5 Sept. 2013), <https://doi.org/10.1093/oxfordhb/9780199579563.013.0063>
- Colombo, M., & Piccinini, G. (2023). *The computational theory of mind*. Cambridge University Press.
- Cominelli, L., Hoegen, G., & De Rossi, D. (2021). Abel: integrating humanoid body, emotions, and time perception to investigate social interaction and human cognition. *Applied Sciences*, 11(3), 1070.
- Copeland, J. (ed.) (2004) *The essential Turing*. Oxford: Oxford University Press.
- Corballis, M. C. (2015). *The wandering mind: What the brain does when you're not looking*. University of Chicago Press.
- Cordeschi, R. (1994). La cibernetica e l'intelligenza artificiale. In *La filosofia degli automi. Origini dell'Intelligenza Artificiale*. Torino: Bollati Boringhieri, 15-52.
- Cordeschi, R. (1998). *La scoperta dell'artificiale: psicologia, filosofia e macchine intorno alla cibernetica*. Bologna: Zanichelli.
- Cordeschi, R. (2004) Cybernetics. In Floridi (Ed.) *The Blackwell Guide to Philosophy of Computing and Information*. p. 186-196
- Cordeschi, R. (2008a). Il metodo sintetico: problemi epistemologici nella scienza cognitiva. *Sistemi intelligenti*, 20 (2), 167-192.

- Cordeschi, R. (2008b). Steps toward the synthetic method. Symbolic information processing and self-organizing systems in early artificial intelligence modeling. In Husbands, P., Holland, O., & Wheeler, M. (Eds.) *The Mechanical Mind in History*. Cambridge, MA: MIT Press.
- Cordeschi, R., & Frixione, M. (2006). Premessa al documento di Dartmouth. *Sistemi intelligenti*, 18(3), 407-413.
- Crane, T. (1999) The Mind-Body Problem. In Rob Wilson & Frank Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.
- Crane, T. (2003) *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. London: Routledge.
- Crombie, A.C. (1961) *Svolte decisive in fisica*. Torino: Bollati Boringhieri
- Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human-machine interactions. *Trends in cognitive sciences*, 25(3), 200-212.
- Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: applying neurocognitive insights to human-robot interaction. *Philosophical Transactions of the Royal Society B*, 374(1771), 20180024.
- Crowell, C. R., Deska, J. C., Villano, M., Zenk, J., & Roddy Jr, J. T. (2019). Anthropomorphism of robots: study of appearance and agency. *JMIR human factors*, 6(2), e12629.
- Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, 107(2), 705-717.
- Cucciniello, I., Sangiovanni, S., Maggi, G., & Rossi, S. (2023). Mind perception in HRI: Exploring users' attribution of mental and emotional states to robots with different behavioural styles. *International Journal of Social Robotics*, 15(5), 867-877.
- Cuffari, E. C., Di Paolo, E., & De Jaegher, H. (2015). From participatory sense-making to language: there and back again. *Phenomenology and the Cognitive Sciences*, 14, 1089-1125.
- Cullen, H., Kanai, R., Bahrami, B., & Rees, G. (2014). Individual differences in anthropomorphic attributions and human brain structure. *Social cognitive and affective neuroscience*, 9(9), 1276-1280.
- Currie, G., & Ravenscroft, I. (2002). *Recreative minds: Imagination in philosophy and psychology*. Clarendon Press.
- Curtoni, V. (2012). *Robot 64*. Milano: Delos Digital srl.
- Dacey, M. (2017). Anthropomorphism as cognitive bias. *Philosophy of Science*, 84(5), 1152-1164.
- Damholdt, M.F., Quick, O.S., Seibt, J. et al. (2023). A Scoping Review of HRI Research on 'Anthropomorphism': Contributions to the Method Debate in HRI. *Int J of Soc Robotics* 15, 1203–1226 (2023). <https://doi.org/10.1007/s12369-023-01014-z>
- Damiano, L. (2010) *Unità in dialogo. Un nuovo stile per la conoscenza*. Milano: Mondadori.

- Damiano, L. (2012). Co-emergences in life and science: a double proposal for biological emergentism. *Synthese*, 185(2), 273-294.
- Damiano, L. (2020). Mente, robot ed ecologie sociali miste. Per un'epistemologia sperimentale dei robot sociali. *Sistemi intelligenti*, 32(1), 27-39.
- Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, 9, 468.
- Damiano, L., & Dumouchel, P. (2020). Emotions in Relation. Epistemological and Ethical Scaffolding for Mixed Human-Robot Social Ecologies. *HUMANA.MENTE Journal of Philosophical Studies*, 13(37), 181-206
- Damiano, L., & Stano, P. (2018). Synthetic Biology and Artificial Intelligence. Grounding a cross-disciplinary approach to the synthetic exploration of (embodied) cognition. *Complex Syst*, 27, 199-228.
- Darling, K. (2021). *The new breed: How to think about robots*. Penguin UK.
- Datteri, E. (2010). Simulazioni, spiegazioni e generalizzazioni nelle scienze cognitive. *Sistemi intelligenti*, 22(3), 459-466.
- Datteri, E. (2013). «I veicoli pensanti» nell'immaginario della robotica contemporanea. *Sistemi intelligenti*, 25(3), 617-636.
- Dautenhahn, K. (1994). Trying to imitate-a step towards releasing robots from social isolation. In *Proceedings of PerAc'94. From Perception to Action* (pp. 290-301). IEEE.
- Dautenhahn, K. (1995). Getting to know each other—artificial social intelligence for autonomous robots. *Robotics and autonomous systems*, 16(2-4), 333-356.
- Dautenhahn, K. (1998a). Embodiment and interaction in socially intelligent life-like agents. In *International Workshop on Computation for Metaphors, Analogy, and Agents* (pp. 102-141). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Dautenhahn, K. (1998b). The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied artificial intelligence*, 12(7-8), 573-617.
- Dautenhahn, K. (2004). Socially Intelligent Agents in Human Primate Culture. In S. Payr & R. Trappl (Eds.), *Agent culture: Human-agent interaction in a multicultural world* (pp. 45–71). Lawrence Erlbaum Associates Publishers.
- Dautenhahn, K. (2007a). Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, 4(1), 15.
- Dautenhahn, K. (2007b). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480), 679-704.
- Dautenhahn, K. (2014). Human-Robot Interaction. *Interaction Design Foundation - IxDF*. <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction>

- Dautenhahn, K., & Nehaniv, C. L. (2002). The agent-based perspective on imitation. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in Animals and Artifacts* (pp. 1–40). MIT Press.
- Dautenhahn, K., & Nehaniv, C. L. (Eds.). (2002). *Imitation in Animals and Artifacts*. MIT Press.
- Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A., & Blow, M. (2009). KASPAR—a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3-4), 369-397.
- Davies, M. & Stone, T. (eds.) (1995). *Folk Psychology: The Theory of Mind Debate*. Blackwell.
- Davis, M. (2000) *The Universal Computer. The Road from Leibniz to Turing*. New York: W. W. Norton & Company (tr. it. *Il calcolatore universale. Da Leibniz a Turing*, Milano, Adelphi, 2003)
- de Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.
- de Graaf, M. M., & Malle, B. F. (2018). People’s judgments of human and robot behaviors: a robust set of behaviors and some discrepancies. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 97-98).
- de Graaf, M. M., & Malle, B. F. (2019). People’s explanations of robot behavior subtly reveal mental state inferences. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 239-248). IEEE.
- De Jaegher, H. (2006). Social interaction rhythm and participatory sense-making. *Unpublished doctoral dissertation*. University of Sussex, Brighton, UK.
- De Jaegher, H. (2009). Social understanding through direct perception? Yes, by interacting. *Consciousness and cognition*, 18(2), 535-542.
- De Jaegher, H. (2018). The intersubjective turn. In Albert Newen, Leon De Bruin, and Shaun Gallagher (eds), *The Oxford Handbook of 4E Cognition*, Oxford Library of Psychology (2018; online edn, Oxford Academic, 9 Oct. 2018), <https://doi.org/10.1093/oxfordhb/9780198735410.013.24>, accessed 15 June 2024.
- De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the cognitive sciences*, 6, 485-507.
- De Jaegher, H., & Di Paolo, E. (2008). Making sense in participation: An enactive approach to social cognition. *Emerging communication*, 10, 33.
- De Jaegher, H., Di Paolo, E. (2007). Participatory sense-making. *Phenom Cogn Sci* 6, 485–507 (2007). <https://doi.org/10.1007/s11097-007-9076-9>
- De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition?. *Trends in cognitive sciences*, 14(10), 441-447.
- de l’Isle-Adam, V. (1886) *Eva Futura*. Venezia: Marsilio.



- De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331.
- de'Sperati, C., & Viviani, P. (1997). The relationship between curvature and velocity in two-dimensional smooth pursuit eye movements. *Journal of Neuroscience*, 17(10), 3932-3945.
- Degenaar, J., & O'Regan, J. K. (2017). Sensorimotor theory and enactivism. *Topoi*, 36, 393-407.
- Dennett, D. C. (1971). Intentional systems. *The journal of philosophy*, 68(4), 87-106.
- Dennett, D. C. (1978a). *Brainstorms*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1978b). Beliefs about beliefs. *The Behavioral and Brain Sciences* 1: 568–570.
- Dennett, D. C. (1978c) Where Am I). In Dennett, D. (1978), *Brainstorms. Philosophical Essays on Mind and Psychology*. Bradford Books. Oggi anche in Hofstadter, D. & Dennett, D. (Eds.) (1981) *The Mind's I. Fantasies and Reflections on Self and Soul*. New York: Basic Books (tr. it. Dove sono? in *L'io della mente*, Milano, Adelphi, 1985: 213-227).
- Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. In Christopher Hookway (ed.), *Minds, Machines and Evolution*. Cambridge, MA: Cambridge University Press.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press. (tr. it. *L'atteggiamento intenzionale*, Il Mulino, Bologna, 1993).
- Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and brain sciences*, 11(3), 495-505.
- Dennett, D. C. (1990). Cognitive wheels: The frame problem of AI. *The philosophy of artificial intelligence*, 147, 170.
- Dennett, D. C. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 349(1689), 133–146. doi:10.1098/rsta.1994.0118
- Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.
- Dennett, D. C. (1997a). Cog as a thought experiment. *Robotics and Autonomous Systems*, 20(2), 251-256.
- Dennett, D. C. (1997b). True believers: The intentional strategy and how it works. In J. Haugeland (Ed.), *Mind design II: Philosophy, psychology, artificial intelligence* (pp. 57-79). MIT Press.
- Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. W. W. Norton & Company. (tr. it. *Dai batteri a Bach: La natura dell'intelligenza* (A. F. Ricci, Trans.), Raffaello Cortina, 2018).
- Derimis, Y., & Hayes, G. (2002). Imitations as a dual-route process featuring predictive and learning components: a biologically plausible computational model. *Imitation in animals and artifacts*, 327-361.

- Descartes, R., *Descartes à Huygens*, in *Œuvres*. Paris: Gallimard, 1949.
- Descartes, R., *Œuvres et lettres*. Paris: Gallimard, 1953.
- Desideri, L., Bonifacci, P., Croati, G., Dalena, A., Gesualdo, M., Molinario, G., Gherardini, A., Cesario, L., & Ottaviani, C. (2021). The mind in the machine: Mind perception modulates gaze aversion during child–robot interaction. *International Journal of Social Robotics*, 13(4), 599–614
- Destephe, M., Brandao, M., Kishi, T., Zecca, M., Hashimoto, K., & Takanishi, A. (2015). Walking in the uncanny valley: Importance of the attractiveness on the acceptance of a robot as a working partner. *Frontiers in psychology*, 6, 204.
- Di Cesare, G., Vannucci, F., Rea, F., Sciutti, A., & Sandini, G. (2020). How attitudes generated by humanoid robots shape human brain activity. *Scientific Reports*, 10(1), 16928.
- Di Dio, C., & Gallese, V. (2009). Neuroaesthetics: a review. *Current Opinion in Neurobiology*, Volume 19, Issue 6, 2009, Pages 682-687, ISSN 0959-4388, <https://doi.org/10.1016/j.conb.2009.09.001>.
- Di Dio, C., Isernia, S., Ceolaro, C., Marchetti, A., & Massaro, D. (2018). Growing up thinking of God's beliefs: Theory of mind and ontological knowledge. *SAGE Open*, 8(4), 2158244018809874.
- Di Dio, C., Manzi, F., Itakura, S., Kanda, T., Ishiguro, H., Massaro, D., & Marchetti, A. (2020). It does not matter who you are: Fairness in pre-schoolers interacting with human and robotic partners. *International Journal of Social Robotics*, 12(5), 1045–1059.
- Di Paolo, E. A. (2003). Organismically-inspired robotics: Homeostatic adaptation and teleology beyond the closed sensorimotor loop. In K. Murase & T. Asakura (Eds.), *Dynamic systems approach for embodiment and sociality: From ecological psychology to robotics* (pp. 19-42).
- Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the cognitive sciences*, 4(4), 429-452.
- Di Paolo, E. A. (2014). The worldly constituents of perceptual presence. *Frontiers in Psychology*, 5, 450.
- Di Paolo, E. A., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic bodies: The continuity between life and language*. MIT press.
- Di Paolo, E. A., Rohde, M., & De Jaegher, H. (2010). Horizons for the enactive mind: Values, social interaction, and play. In Stewart, J., Gapenne, O. & Di Paolo, E. (Eds.) *Enaction: Towards a New Paradigm for Cognitive Science*. Cambridge (MA): MIT Press.
- Di Paolo, E., & Thompson, E. (2014). The enactive approach. In *The Routledge handbook of embodied cognition* (pp. 68-78). Routledge.
- Di Paolo, E., Beer, R. D., & Thompson, E. (2021). Laying down a forking path: Incompatibilities between enaction and the free energy principle. <https://doi.org/10.31234/osf.io/d9v8f>
- Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.



- Di Paolo, E., Rohde, M., & De Jaegher, H. (2007). Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart, O. Gapenne, & E. Di Paolo (Eds.), *Enaction: Towards a new paradigm for cognitive science* (pp. 33-87). MIT Press.
- Di Paolo, E., Rohde, M., & De Jaegher, H. (2018). *Linguistic bodies: The continuity between life and language*. MIT Press.
- Diderot, D., & d'Alembert, J. L. R. (1751-1772). *Encyclopédie, ou, Dictionnaire raisonné des sciences, des arts et des métiers*. Pergamon Press.
- Doncieux, S., Bredeche, N., Mouret, J. B., & Eiben, A. E. (2015). Evolutionary robotics: what, why, and where to. *Frontiers in Robotics and AI*, 2, 4.
- Draude, C. (2011). Intermediaries: reflections on virtual humans, gender, and the Uncanny Valley. *AI & society*, 26, 319-327.
- Dreyfus, H. L. (1965) *Alchemy and Artificial Intelligence*. Technical Report, Research Report, RAND CORP SANTA MONICA CA.
- Dreyfus, H. L. (1967) Why computers must have bodies in order to be intelligent. *Review of Metaphysics*, 21: 13-32.
- Dreyfus, H. L. (1972) *What Computers Can't Do. A Critique of Artificial Reason*. New York: Harper & Row.
- Dreyfus, H. L. (1991), *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. Cambridge, MA: MIT Press.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT press.
- Dreyfus, H. L. (2002). Intelligence without representation: The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences*, 1(4), 367-383.
- Dreyfus, H. L. (2007). The Return of the Myth of the Mental. *Inquiry*, 50(4), 352-365.
- Dreyfus, H. L. (2008) Why Heideggerian AI failed and How Fixing It Would Require Making It More Heideggerian. In Husbands, P., Holland, O., & Wheeler, M. (Eds.) *The Mechanical Mind in History*. Cambridge (MA): MIT Press, 331-371.
- Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine*. New York: Simon and Schuster.
- Dreyfus, H. L., & Dreyfus, S. E. (1988). Making a mind versus modelling the brain: Artificial intelligence back at the branchpoint. *Artificial Intelligence* 117, 15-44.
- Drummond, J. (2012). Intentionality without representationalism. In D. Zahavi (Ed.), *The Oxford handbook of contemporary phenomenology* (pp. 245-260). Oxford University Press.
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3-4), 177-190.
- Dumouchel, P., & Damiano, L. (2011). Artificial empathy, imitation and mimesis. *Ars Vivendi Journal*, 1, 18-31.

- Dumouchel, P., & Damiano, L. (2017). *Living with robots*. Harvard University Press. (tr. it. *Vivere con il robot. Saggio sull'empatia artificiale*, Milano, Cortina, 2019).
- Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 6(5), 178-190.
- Dunbar, R. I. (2023). The Social Brain Hypothesis Thirty Years On: Some Philosophical Pitfalls of Deconstructing Dunbar's Number. In *Annales Academiae Scientiarum Fennicae* (No. 1, pp. 8-27).
- Dupuy, J. P. (1994) *Aux origines des sciences cognitives*. Paris: Éditions La Découverte (tr. it. *Alle origini delle scienze cognitive. La meccanizzazione della mente*, Milano, Mimesis, 2014).
- Duranti, A. (1999). Intentionality. *Journal of Linguistic Anthropology*, 9(1/2), 134-136.
- Duval, C., Piolino, P., Bejanin, A., Eustache, F., & Desgranges, B. (2011). Age effects on different components of theory of mind. *Consciousness and Cognition*, 20(3), 627-642. <https://doi.org/10.1016/j.concog.2010.10.025>
- Eco, U. (1984). *Semiotica e filosofia del linguaggio*. Milano: Bompiani.
- Eco, U. (1993) L'«ars magna» di Raimondo Lullo. In Eco, U. *La ricerca della lingua perfetta*. Roma-Bari: Laterza.
- Eco, U. (1997). *Kant e l'ornitorinco*. Milano: Bompiani.
- Ellis, L. U., Sims, V. K., Chin, M. G., Pepe, A. A., Owens, C. W., Dolezal, M. J., ... & Finkelstein, N. (2005). Those a-maze-ing robots: Attributions of ability are based on form, not behavior. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 3, pp. 598-601). Sage CA: Los Angeles, CA: SAGE Publications.
- Epley, N. (2018). A mind like mine: The exceptionally ordinary underpinnings of anthropomorphism. *Journal of the Association for Consumer Research*, 3(4), 591-598.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4), 864.
- Erden, M. S. (2013). Emotional postures for the humanoid-robot nao. *International Journal of Social Robotics*, 5, 441-456.
- Essaka, J. (2001). *Pygmalion and Galatea: The History of a Narrative in English Literature*. Ashgate.
- Esterwood, C., & Robert, L. P. (2023). The theory of mind and human–robot trust repair. *Scientific Reports*, 13(1), 9877.
- Evans, J. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454–459.
- Eyssel, F., & Kuchenbrandt, D. (2012). "Social categorization of social robots: Anthropomorphism as a function of robot group membership." *British Journal of Social Psychology*, 51(4), 724-731.

- Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. In *2013 8th acm/ieee international conference on human-robot interaction (hri)* (pp. 121-122). IEEE.
- Fabbris, L. (2021) Critica della ragion pregettuale: storia di un cervello senza progetto. Introduzione a: Ashby, W. R., *Progetto per un cervello. L'origine del comportamento adattivo*. Napoli-Salerno, Orthotes, 2021, 7-44.
- Falck-Ytter, T., Gredebäck, G., & Von Hofsten, C. (2006). Infants predict other people's action goals. *Nature neuroscience*, 9(7), 878-879.
- Feigenbaum, E., & Shrobe, H. (1993). The Japanese national Fifth Generation project: introduction, survey, and evaluation. *Future Generation Computer Systems*, 9(2), 105-117
- Feng, S., Wang, X., Wang, Q., Fang, J., Wu, Y., Yi, L., & Wei, K. (2018). The uncanny valley effect in typically developing children and its absence in children with autism spectrum disorders. *PloS one*, 13(11), e0206343.
- Ferrari, P. F., & Gallese, V. (2007). Mirror neurons and intersubjectivity. *On being moved*, 73-88.
- Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. *Frontiers in psychology*, 6, 249.
- Fiala, B., Arico, A., & Nichols, S. (2014). You, robot. In E. Machery & E. O'Neill (Eds.), *Proceedings of the Current Controversies in Experimental Philosophy* (pp. 31–47). Routledge.
- Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4* (pp. 199-208). Springer Berlin Heidelberg.
- Fischer, K. (2021). Tracking anthropomorphizing behavior in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(1), 1-28.
- Fisher, J. A. (1996). The Myth of Anthropomorphism John Andrew Fisher. In Dale Jamieson & Marc Bekoff (eds.), *Readings in Animal Cognition*. MIT Press.
- Flach, R., Knoblich, G., & Prinz, W. (2004). The two-thirds power law in motion perception. *Visual Cognition*, 11(4), 461-481.
- Flamm, C. (2019). *Galatea: A Representation of the Nature of the Goddess*. A Thesis Submitted to the Faculty of the Art and Design Department in Partial Fulfillment of the Requirements for the Degree of Master of Arts in Art History at Lindenwood University.
- Flash, T., & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience*, 5(7), 1688-1703.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28, 689-707.
- Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
- Fodor, J. A. (1981). The mind-body problem. *Scientific american*, 244(1), 114-123.

- Fodor, J. A. (1983). *The modularity of mind*. MIT press.
- Fodor, J. A. (2008). Against Darwinism. *Mind & Language* 23: 1-24.
- Fodor, J. A. & Lepore, E. (1992). *Holism: A Shopper's Guide*. Cambridge: Blackwell
- Foehr, J. A. & Germelmann, C. C. (2020). Alexa, can I trust you? Exploring consumer paths to trust in smart voice-interaction technologies. *Journal of the Association for Consumer Research*, 5(2), 181-205.
- Foerster, H. von (1950-5) *Cybernetics, Circular Causal, and Feedback Mechanisms in Biological and Social Systems: Published Transactions of the Sixth, Seventh, Eighth, Ninth, and Tenth Conference*, 5 vols. New York: Josiah Macy Foundaton.
- Foerster, H. von (1960) On Self-Organizing Systems and their Environments. In Yovits, M. (Ed.), *Self-Organizing Systems*. London: Pergamon, 30-50.
- Foerster, H. von (1962) Perception of Form in Biological and Man-Made Systems. *Trans. Of the Indust. Design Educ. Assoc. (IDEA) Symp.* E.J. Zagorski (Ed.) University of Illinois, Urbana, 10-37.
- Foerster, H. von (1987). *Sistemi che osservano*. Ceruti, M. & Telfner, U. (Eds.). Roma: Astrolabio.
- Foerster, H., von (2003). *Understanding Understanding: Essays on Cybernetics and Cognition*. New York: Springer-Verlag.
- Foerster, H., von et al. (Eds) (1974), *Cybernetics of Cybernetics, BCL Report 73.38*, Biological Computer Laboratory, Department of Electrical Engineering, University of Illinois, Urbana, IL.
- Fogel, L. J., McCulloch, W. S., & Ramsey-Klee, D. (1970). Natural automata and prosthetic devices. *Aids to biological communication: Prothesis and synthesis*, 2, 221-262.
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4), 143-166.
- Ford, K. M., & Pylyshyn, Z. W. (Eds.) (1996). *The robot's dilemma revisited: The frame problem in artificial intelligence*. Ablex Publishing.
- Fortino, A., & Gaur, S. S. (2024) Generative Artificial Intelligence as the Killer App: Revolutionizing Human–Robot Interaction. In *Synergy of AI and Fintech in the Digital Gig Economy* (pp. 357-369). CRC Press.
- Fortunati, L. (2000). Mostro. In *Enciclopedia Italiana Treccani - Universo del Corpo*: [https://www.treccani.it/enciclopedia/mostro\\_\(Universo-del-Corpo\)/](https://www.treccani.it/enciclopedia/mostro_(Universo-del-Corpo)/)
- Fortunati, L., Manganelli, A.M., Höflich, J. et al. (2023). Exploring the Perceptions of Cognitive and Affective Capabilities of Four, Real, Physical Robots with a Decreasing Degree of Morphological Human Likeness. *Int J of Soc Robotics* 15, 547–561 (2023). <https://doi.org/10.1007/s12369-021-00827-0>
- Frankenhuis, W. E., Barrett, H. C., & Johnson, S. P. (2013). Developmental origins of biological motion perception. *People watching: Social, perceptual, and neurophysiological studies of body perception*, 123-140.

- French, R. M. (2000). The Turing Test: the first 50 years. *Trends in cognitive sciences*, 4(3), 115-122.
- Freud, S. (1919). Il perturbante. In *Sigmund Freud. Saggi sull'arte, la letteratura e il linguaggio*, 1969, Bollati, Torini, pp. 267-307.
- Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456), 815-836.
- Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, 11(2), 127-138.
- Frith, C. (1994). Theory of mind in schizophrenia. In A. S. David & J. C. Cutting (Eds.), *The neuropsychology of schizophrenia* (pp. 147–161). Lawrence Erlbaum Associates, Inc.
- Frith, C. (2009). Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3453-3458.
- Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, 15(17), R644-R645. <https://doi.org/10.1016/j.cub.2005.08.041>
- Frith, C., & Frith, U. (2006). How we predict what other people are going to do. *Brain research*, 1079(1), 36-46.
- Froese, T. (2010). From cybernetics to second-order cybernetics: a comparative analysis of their central ideas. *Constructivist Foundations*, 5(2).
- Froese, T., & Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics & Cognition*, 19(1), 1-36.
- Froese, T., & Gallagher, S. (2012). Getting interaction theory (IT) together: Integrating developmental, phenomenological, enactive, and dynamical approaches to social interaction. *Interaction Studies*, 13(3), 436-468.
- Froese, T., & Taguchi, S. (2019). The problem of meaning in AI and robotics: Still with us after all these years. *Philosophies*, 4(2), 14.
- Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial intelligence*, 173(3-4), 466-500.
- Fuchs, T. (2017). The “As-If” Function and Its Loss in Schizophrenia 1. In *Imagination and social perspectives* (pp. 83-98). Routledge.
- Fuchs, T. (2019). The uncanny as atmosphere. *Psychopathology and Atmospheres: Neither inside nor Outside*, Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 101-118.
- Fuchs, T. (2021). *In defense of the human being: Foundational questions of an embodied anthropology*. Oxford University Press.
- Fuchs, T. (2024). Understanding Sophia? On human interaction with artificial agents. *Phenomenology and the Cognitive Sciences*, 23(1), 21-42.
- Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the cognitive sciences*, 8, 465-486.

- Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008, March). How people anthropomorphize robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (pp. 145-152).
- Gahrn-Andersen, R. (2022). Seeming autonomy, technology and the uncanny valley. *AI & society*, 37(2), 595-603.
- Gallagher, S. (2001). The practice of mind. Theory, simulation or primary interaction?. *Journal of consciousness studies*, 8(5-6), 83-108.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford Academic.
- Gallagher, S. (2007). Social cognition and social robots. *Pragmatics & Cognition*, 15(3), 435-453.
- Gallagher, S. (2008a). Inference or interaction: Social cognition without precursors. *Philosophical Explorations*, 11(3), 163-174.
- Gallagher, S. (2008b). Direct perception in the intersubjective context. *Consciousness and cognition*, 17(2), 535-543.
- Gallagher, S. (2011). Interpretations of embodied cognition. In W. Tschacher and C. Bergomi (eds.), *The Implications of Embodiment: Cognition and Communication*. Exeter: Imprint Academic.
- Gallagher, S. (2012a). Neurons, neonates and narrative: From embodied resonance to empathic understanding. In A. Foolen, U. Lüdtke, J. Zlatev and T. Racine (eds.), *Moving Ourselves, Moving Others* (167-96). Amsterdam: John Benjamins.
- Gallagher, S. (2012b). Social cognition, the Chinese room, and the robot replies. In *Knowing without thinking: Mind, action, cognition and the phenomenon of the background* (pp. 83-97). London: Palgrave Macmillan UK.
- Gallagher, S. (2013a). When the problem of intersubjectivity becomes the solution. *The infant mind: Origins of the social brain*, 48-74.
- Gallagher, S. (2013b). You and I, robot. *AI & society*, 28, 455-460.
- Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.
- Gallagher, S. (2020a). *Action and interaction*. Oxford University Press.
- Gallagher, S. (2020b). The brains behind radical ecological and enactive approaches to cognition. *Balzan papers: III, 2020.-(Balzan papers; 3)*, 355-367.
- Gallagher, S. (2023a). *Embodied and enactive approaches to cognition*. Cambridge University Press.
- Gallagher, S. (2023b). Surprise! Why enactivism and predictive processing are parting ways: The case of improvisation. *Possibility Studies & Society*, 1(3), 269-278.
- Gallagher, S. & Bower, M. (2013). Making enactivism even more embodied. *AVANT. Pismo Awangardy Filozoficzno-Naukowej*, 2, 232-247.
- Gallagher, S. & Zahavi, D. (2008). *The Phenomenological Mind*. New York, NY: Routledge. Edited by Dan Zahavi (tr. it. *La mente fenomenologica*, Milano, Cortina, 2022).

- Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 195(6), 2627-2648.
- Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. *The shared mind: Perspectives on intersubjectivity*, 12, 17-38.
- Gallagher, S., & Miyahara, K. (2012). Neo-pragmatism and enactive intentionality. In J. Schulkin (Ed.), *Action, perception and the brain: New directions in philosophy and cognitive science* (pp. 97–113). Palgrave Macmillan.
- Gallagher, S., Hutto, D., & Hipólito, I. (2022). Predictive processing and some disillusiones about illusions. *Review of Philosophy and Psychology*, 13(4), 999-1017.
- Gallese, V. (2020). Brain, Body, Habit and the Performative Quality of Aesthetics. In I. Testa & F. Caurana (editors), *Habits: Pragmatist Approaches from Cognitive Neuroscience to Social Science*. Cambridge, UK: Cambridge University Press.
- Gallese, V. (2024). Digital visions: the experience of self and others in the age of the digital revolution. *International Review of Psychiatry*, 1-11.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12), 493-501.
- Gallese, V., Rochat, M., Cossu, G., & Sinigaglia, C. (2009). Motor cognition and its role in the phylogeny and ontogeny of action understanding. *Developmental psychology*, 45(1), 103.
- Gangopadhyay, N., & Kiverstein, J. (2009). Enactivism and the unity of perception and action. *Topoi*, 28, 63-73.
- Gargano, A., Cominelli, L., Vannucci, C., Cecchetti, L., & Scilingo, E. P. (2022, October). Preliminary personality model for social robots based on the Cognitive-Affective Processing System theory. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)* (pp. 223-228). IEEE.
- Gärtner, K., & Clowes, R. W. (2017). Enactivism, Radical Enactivism and Predictive Processing: What is Radical in Cognitive Science?. *Kairos. Journal of Philosophy & Science*, 18(1), 54-83.
- Gazzola, V. et al. (2007) The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage* 35, 1674–1684
- Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56(2), 165-193.
- Gerrans, P. (2009). Imitation and theory of mind. *Handbook of neuroscience for the behavioral sciences*, 2, 905-922.
- Geue, L. (2021). *From Robots to Primates: Tracing the Uncanny Valley Effect to its Evolutionary Origin* (Bachelor's thesis, University of Twente).
- Geymonat, L. (1976). Trasformazioni di fondo della scienza fisica. In *Storia del pensiero filosofico e scientifico*, vol. 8, cap XIII, pp. 427-444.

- Ghiglino, D., Marchesi, S., & Wykowska, A. (2023). Play with me: complexity of human-robot interaction affects individuals' variability in intentionality attribution towards robots.
- Gigerenzer, G., & Goldstein, D. G. (1996). Mind as computer: Birth of a metaphor. *Creativity Research Journal*, 9(2-3), 131-144.
- Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., Guazzelli, M., & Pietrini, P. (2011). Distinct neural systems involved in agency and animacy detection. *Journal of Cognitive Neuroscience*, 23(8), 1911–1920.
- Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1-48.
- Goertzel, B., Lian, R., Arel, I., de Garis, H., Chen, S. (2010) A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. *Neurocomputing*, Volume 74, Issues 1–3, 2010, Pages 30-49.
- Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.
- Goodrich, M. A. (2004, October). Using Models of Cognition in HRI Evaluation and Design. In *AAAI Technical Report (5)* (pp. 17-24).
- Goodrich, M. A., & Schultz, A. C. (2008). Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3), 203-275.
- Gopnick, A. & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, Vol. 7, No. ½, pp. 145-171.
- Gopnik, A. (1996). The scientist as child. *Philosophy of science*, 63(4), 485-514.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257-293). New York: Cambridge University Press.
- Gordon, R. M., & Cruz, J. (2003). Simulation theory. *Encyclopedia of cognitive science*, 4, 9-14.
- Gozzano, S. (1997). *Storia e teorie dell'intenzionalità. Il dibattito contemporaneo e le sue radici*. Roma-Bari: Laterza.
- Graf, A. (1996) *Miti, leggende e superstizioni del Medio Evo*. Milano: Mondadori.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619-619.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *science*, 315(5812), 619-619.
- Gray, H., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619.



- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125-130.
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125-130.
- Green, R. D., MacDorman, K. F., Ho, C. C., & Vasudevan, S. (2008). Sensitivity to the proportions of faces that vary in human likeness. *Computers in Human Behavior*, 24(5), 2456-2474.
- Grey Walter, W. (1950a). An imitation of life. *Scientific American*, 182(5), 42-45.
- Grey Walter, W. (1950b). An electro-mechanical animal. *Dialectica*, 4(3), 206-213.
- Grey Walter, W. (1951). A machine that learns. *Scientific American*, 184(8), 60-63.
- Grey Walter, W. (1963). *The Living Brain*. New York, NY: W.W. Norton & Co.
- Grigore, E. C., Pereira, A., Zhou, I., Wang, D., & Scassellati, B. (2016). Talk to me: Verbal communication improves perceptions of friendship and social presence in human-robot interaction. In *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16* (pp. 51-63). Springer International Publishing.
- Gunes, H., Broz, F., Crawford, C. S., der Pütten, A. R. V., Strait, M., & Riek, L. (2022). Reproducibility in human-robot interaction: furthering the science of HRI. *Current Robotics Reports*, 3(4), 281-292.
- Gunkel, D. J. (2012). *The Machine Question*. MIT Press.
- Gunkel, D. J. (2014). A Vindication of the Rights of Machines. *Philosophy & Technology*, 27, 113-132.
- Guthrie, S. & Porubanova, M. (2022). Faces in Clouds and Voices in the Wind: Anthropomorphism, Agency Detection and Human Cognition. In *The Routledge Handbook of Evolutionary Approaches to Religion* (pp. 94-110)
- Guthrie, S. E. (1995). *Faces in the clouds: A new theory of religion*. Oxford University Press.
- Guthrie, S., & Porubanova, M. (2020). Faces in clouds and voices in wind: Anthropomorphism in religion and human cognition. *Routledge handbook of evolutionary approaches to religion*. Routledge.
- Hacking, I. (1989) Putnam's Change of Mind. *London Review of Books*, Vol. 11 No. 9
- Hacking, I. (2005). The Cartesian vision fulfilled: analogue bodies and digital minds. *Interdisciplinary Science Reviews*, 30(2), 153-166.
- Hakli, R., & Seibt, J. (2017). *Sociality and normativity for robots*. Cham: Springer.
- Hallowell, R. (2009). Humberto Maturana and Francisco Varela's contribution to media ecology: Autopoiesis, the Santiago School of Cognition, and enactive cognitive science. *Proceedings of the Media Ecology Association*, 10, 143-158.

- Hanna, R., & Thompson, E. (2003). The mind-body-body problem. *Theoria et Historia Scientiarum*, 7(1), 23-42.
- Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., ... & Stephanou, H. (2005). Upending the uncanny valley. In *AAAI* (Vol. 5, pp. 1728-1729).
- Hao, K. (2023, May 2). *Geoffrey Hinton tells us why he's now scared of the tech he helped build*. MIT Technology Review. Disponibile all'indirizzo: <https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/>
- Happé, F., & Frith, U. (1995). Theory of mind in autism. In *Learning and cognition in autism* (pp. 177-197). Boston, MA: Springer US.
- Haring, K. S., Silvera-Tawil, D., Matsumoto, Y., Velonaki, M., & Watanabe, K. (2013). Perception of an android robot in Japan and Australia: A cross-cultural comparison. *Proceedings of the 8th ACM/IEEE international conference on Human-Robot Interaction (HRI)*, 291-292.
- Haring, K. S., Silvera-Tawil, D., Takahashi, T., Velonaki, M., & Watanabe, K. (2015). Perception of a humanoid robot: A cross-cultural comparison. In *Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 821-826).
- Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1), 5-25.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.
- Harney, M. J. (2013). *Intentionality, sense and the mind*. Springer Science & Business Media.
- Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2), 100-107.
- Haslam, M. (2023). Anthropomorphism as a contributor to the success of human (*Homo sapiens*) tool use. *Journal of Comparative Psychology*, 137(3), 200.
- Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, 10(3), 252-264.
- Haslam, N., Loughnan, S., Reynolds, C., & Wilson, S. (2007). Dehumanization: A new perspective. *Social and Personality Psychology Compass*, 1(1), 409-422.
- Haugeland, J. (1989). *Artificial intelligence: The very idea*. Cambridge, MA: MIT press.
- Haugeland, J. (1990). The intentionality all-stars. *Philosophical perspectives*, 4, 383-427.
- Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008b). Theory of mind (ToM) on robots: A functional neuroimaging study. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (pp. 335-342).
- Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008a). Understanding social robots: A user study on anthropomorphism. In *RO-MAN 2008-the 17th IEEE international symposium on robot and human interactive communication* (pp. 574-579). IEEE.

- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2), 243-259.
- Heisenberg, W. (1958) *Physics and Philosophy. The Revolution in Modern Science*. New York: Harper & Brothers.
- Helmholtz, H. von (1854). *Über die Wechselwirkung der Naturkräfte. Ein populär-wissenschaftlicher Vortrag, gehalten am 7. Februar 1854 in Königsberg in Preussen*, in *Populäre wissenschaftliche Vorträge*, 2. Heft, Vieweg, Braunschweig 1871, pp. 139 sgg. Traduzione in Losano 1990, p. 136-137.
- Henderson, H. (2009). *Encyclopedia of computer science and technology*. Infobase Publishing.
- Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences*, 43(6), 373-384.
- Heppenheimer, T. (1985) *L'uomo fatto dall'uomo*. In Minsky, M. (Ed.) (1985) *Robotics: The First Authoritative Report from the Ultimate High-Tech Frontier*. New York Doubleday. (tr. it. Sosio, L., *La robotica. Il primo autorevole rapporto delle frontiere dell'alta tecnologia*. Milano: Longanesi, 1987, 25-70).
- Hermans, C. (2023). Of rhythm and movement: physical play and dance as (participatory) sense-making practices. *Research in Dance Education*, 1–16. <https://doi.org/10.1080/14647893.2023.2211524>
- Hernández-Orallo, J. (2020). Twenty years beyond the Turing test: moving beyond the human judges too. *Minds and Machines*, 30(4), 533-562.
- Hersey, G. L. (2009). *Falling in love with statues: artificial humans from Pygmalion to the present*. University of Chicago Press.
- Hertzberg, J. (2007) Embodying AI — GOF AI Goes Robotics. *IK-2007*.
- Hicheur, H., Vieilledent, S., Richardson, M.J.E. *et al.* Velocity and curvature in human locomotion along complex curved paths: a comparison with hand movements. *Exp Brain Res* **162**, 145–154 (2005). <https://doi.org/10.1007/s00221-004-2122-8>
- Hilton, D. (2012). The emergence of cognitive social psychology: A historical analysis. In *Handbook of the history of social psychology* (pp. 45-79). Psychology Press.
- Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human–Computer Interaction*, 19(1-2), 151-181.
- Hipólito, I., Hutto, D. D., & Gallagher, S. (2020). Culture in mind-an enactivist account: Not cognitive penetration but cultural permeation. In Laurence J. Kirmayer, Carol M. Worthman, Shinobu Kitayama, Robert Lemelson & Constance Cummings (eds.), *Culture, Mind, and Brain: Emerging Concepts, Models, and Applications*. Cambridge University Press.
- Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, 26(6), 1508-1518

- Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *International Journal of Social Robotics*, 9, 129-139.
- Ho, C.-C., MacDorman, K. F., & Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and Isomap analysis of robot video ratings. In *Proceedings of the third ACM/IEEE international conference on human-robot interaction* (pp. 169–176). <http://dx.doi.org/10.1145/1349822.1349845>
- Hoffmann, E. T. A. (1962). Der Sandmann [1817]. *ETAH, Sämtliche Werke in sechs Bänden*, 3, 1816-1820.
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Basic Books (tr. it. *Gödel, Escher, Bach: un'Eterna Ghirlanda Brillante*, Milano, Adelphi, 1984).
- Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic Books (tr. it. *Concetti fluidi e analogie creative*. Milano: Adelphi, 1996)
- Hogan, K. (2017). Is the machine question the same question as the animal question?. *Ethics and Information Technology*, 19(1), 29-38.
- Hohwy, J. (2013). *The predictive mind*. OUP Oxford.
- Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, 35(2), 209-223.
- Holland, O. (1997). “Grey Walter: the pioneer of real artificial life,” in *Proceedings of the 5th International Workshop on Artificial Life*, 34–44. Cambridge: MIT Press.
- Horáková, J., & Kelemen, J. (2008). The robot story: why robots were born and how they grew up. In Husbands, P., Holland, O., & Wheeler, M. (Eds.) *The Mechanical Mind in History*. Cambridge (MA): MIT Press, 283-306.
- Horgan, T., & Kriegel, U. (2008). Phenomenal intentionality meets the extended mind. *The Monist*, 91(2), 347-373.
- Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: the scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences*, 1426(1), 93-110.
- Hsu, H. (2023). Oral History of Rodney Brooks. *IEEE Annals of the History of Computing*, 45(4), 92-103.
- Hsu, J. (2012, June 12). Robotics’ Uncanny Valley gets new translation. *Livescience*. Disponibile in: <https://www.livescience.com/20909-robotics-uncanny-valley-translation.html>
- Huang, K. H., & Deng, Y. S. (2008). Social interaction design in cultural context: A case study of a traditional social activity. *International journal of Design*, 2(2).
- Husbands, P. (2014) Robotics. In Frankish, K. & Ramsey, W. (Eds.) *The Cambridge Handbook of Artificial Intelligence*, Cambridge, MA: Cambridge University Press, 269-295.

- Hutto, D. D. (2006). Embodied expectations and extended possibilities: Reply to Goldie. In Menary, R. (Ed.) *Radical Enactivism: Intentionality, Phenomenology and Narrative*. John Benjamin.
- Hutto, D. D. (2007). The narrative practice hypothesis: Origins and applications of folk psychology. *Royal Institute of Philosophy Supplements*, 60, 43-68.
- Hutto, D. D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons*. Cambridge (MA): MIT press.
- Hutto, D. D. (2009). Folk psychology as narrative practice. *Journal of Consciousness Studies*, 16(6-7), 9-39.
- Hutto, D. D. (2015). Overly enactive imagination? Radically re-imagining imagining. *The Southern Journal of Philosophy*, 53, 68-89.
- Hutto, D. D., & Hipólito, I. (2024). Updating our theories of perceiving: from predictive processing to radical enactivism. In *The Roles of Representation in Visual Perception* (pp. 441-461). Cham: Springer International Publishing.
- Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds meet content*. MIT Press
- Hutto, D. D., & Myin, E. (2017). *Enaction: A new paradigm for cognitive science*. MIT Press.
- Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia*, 43, 521-536.
- Hutto, D. D., Herschbach, M., & Southgate, V. (2011). Social cognition: mindreading and alternatives. *Review of Philosophy and Psychology*, 2, 375-395.
- Hutto, D., & Ravenscroft, I. (2021). Folk psychology as a theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). Retrieved from <https://plato.stanford.edu/archives/fall2021/entries/folkpsych-theory/>
- Iacoboni, M. (2009). *Mirroring people: The new science of how we connect with others*. Farrar, Straus and Giroux.
- Iliade, Libro XVIII. Traduzione di Monti, V., Milano, 1978, in Cohen, J. (1966).
- Imamura, Y., Terada, K., & Takahashi, H. (2015). Effects of behavioral complexity on intention attribution to robots. In *Proceedings of the 2015 3rd International Conference on Human-Agent Interaction* (pp. 65–72).
- Isakowitsch, C. (2022). How augmented reality beauty filters can affect self-perception. In *Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 239-250). Cham: Springer Nature Switzerland.
- Ishiguro, H. (2006a). Android science: Toward a new cross-interdisciplinary framework. In *Robotics research: results of the 12th International Symposium ISRR* (pp. 118-127). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Ishiguro, H. (2006b). Interactive humanoids and androids as ideal interfaces for humans. In *Proceedings of the 11th international conference on Intelligent user interfaces* (pp. 2-9).

- Ishiguro, H. (2007). Scientific issues concerning androids. *The International Journal of Robotics Research*, 26(1), 105-117.
- Ishiguro, H. (2020). *How human is human?: The view from Robotics Research*. JPIC (tr. it. *Come costruire un essere umano*, WUDZ, 2024).
- Ishiguro, H., & Dalla Libera, F. (Eds.) (2018). *Geminoid studies: Science and technologies for humanlike teleoperated androids*. Springer.
- Ishihara, S., Nagamachi, M., Schütte, S., & Eklund, J. (2008). Affective meaning: The kansei engineering approach. In *Product experience* (pp. 477-496). Elsevier.
- Israel, G. (2004). *La macchina vivente. Contro le visioni meccanicistiche dell'uomo*. Torino: Bollati Boringhieri.
- Israel, G. (2015). *Meccanicismo. Trionfi e miserie della visione meccanica del mondo*. Bologna: Zanichelli.
- Ito, A., Hayakawa, S., & Terada, T. (2004). Why robots need body for mind communication: An attempt of eye-contact between human and robot. In *Proceedings of the 2004 13th IEEE International Workshop on Robot and Human Interactive Communication* (pp. 473–478)
- Ito, T. A. & Urland, G. R. (2005). The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective, & Behavioural Neuroscience*, 5, 21–36
- Jack, A. I., Dawson, A. J., & Norr, M. E. (2013a). Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization. *NeuroImage*, 79, 313-328.
- Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccio, A. H., & Snyder, A. Z. (2013b). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, 66, 385-401.
- Jacob, P. (2003). Intentionality. In *The Stanford Encyclopedia of Philosophy* (Fall 2003 Edition). Retrieved from <https://plato.stanford.edu/archives/fall2003/entries/intentionality/>.
- Jaynes, J. (1976). *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston: Houghton Mifflin (tr. it. *Il crollo della mente bicamerale e l'origine della coscienza*. Adelphi, Milano, 1984).
- Jentsch, E. (1906/1997). Jentsch, E. (1997). On the psychology of the uncanny. *Angelaki*, 2(1), 7–16. <https://doi.org/10.1080/09697259708571910>
- Ji, Y., Yang, Y., Shen, F., Shen, H. T., & Li, X. (2019). A survey of human action analysis in HRI applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7), 2114-2128.
- Jo, A. (2023). The promise and peril of generative AI. *Nature*, 614(1), 214-216.
- Johanson, D. L., Ahn, H. S., & Broadbent, E. (2021). Improving interactions with healthcare robots: a review of communication behaviours in social and healthcare contexts. *International Journal of Social Robotics*, 13(8), 1835-1850.

- Johnston, J. L., Kearney, G. P., Gormley, G. J., & Reid, H. (2020). Into the uncanny valley: Simulation versus simulacrum?. *Medical Education*, 54(10), 903-907.
- Jost, C., Le Pévédic, B., Belpaeme, T., Bethel, C., Chrysostomou, D., Crook, N., Grandgeorge, M., Mirnig, N. (Eds.) (2020) *Human-Robot Interaction. Evaluation Methods and Their Standardization*. Springer, Cham.
- Kaelbling, L. P. (1991). A situated-automata approach to the design of embedded agents. *ACM SIGART Bulletin*, 2(4), 85-88.
- Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... & Shen, S. (2012). “Robovie, you’ll have to go into the closet now”: Children’s social and moral relationships with a humanoid robot. *Developmental psychology*, 48(2), 303.
- Kahn, K. B., Goff, P. A., & McMahon, J. M. (2015). Intersections of Prejudice and Dehumanization. In Wulf D. Hund, Charles W. Mills, Silvia (Eds.) *Simianization: Apes, Gender, Class, and Race*, Vienna: LIT.
- Kahn, P. H., Jr., Friedman, B., Pe´rez-Granados, D. R., & Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies*, 7, 405–436. doi:10.1.1.130.3260
- Kahn, P. H., Jr., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., ... & Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 159-160).
- Kamide, H., Eyssel, F., & Arai, T. (2013). Psychological anthropomorphism of robots. In *Proceedings of the International Conference on Social Robotics* (pp. 199–208). Springer
- Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature Communications*, 2, 341.
- Kanda, T. & Ishiguro, H. (2013). *Human–Robot Interaction in Social Robotics. Volume 1*. Boca Raton, FL: CRC Press.
- Kanda, T., & Ishiguro, H. (2017). *Human-robot Interaction in Social Robotics. Volume 2*. Boca Raton, FL: CRC Press.
- Kandel, S., Orliaguet, J. P., & Viviani, P. (2000). Perceptual anticipation in handwriting: The role of implicit motor competence. *Perception & Psychophysics*, 62, 706-716.
- Kang, M. (2011). *Sublime dreams of living machines*. Cambridge, MA: Harvard University Press.
- Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International journal of humanoid robotics*, 1(03), 465-480.
- Katara, P., Xian, Z., & Fragkiadaki, K. (2024, May). Gen2sim: Scaling up robot learning in simulation with generative models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6672-6679). IEEE.
- Katič, D., & Vukobratović, M. (1996). Connectionist based robot control: an overview. *IFAC Proceedings Volumes*, 29(1), 169-174.



- Kätsyri, J., Förger, K., Mäkräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in psychology*, 6, 390.
- Keijsers, M., Bartneck, C., & Eyssel, F. (2022). Pay them no mind: the influence of implicit and explicit robot mind perception on the right to be protected. *International Journal of Social Robotics*, 14(2), 499-514.
- Keltner, D., Ekman, P., Gonzaga, G. C. & Beer, J. (2003). Facial expression of emotion. In: R. J. Davidson, K. R. Scherer & H. H. Goldsmith, eds. *Handbook of Affective Sciences*, pp. 415–432. Oxford: Oxford University Press
- Kidd, C. D., & Breazeal, C. (2004, September). Effect of a robot on user perceptions. In *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (IEEE Cat. No. 04CH37566) (Vol. 4, pp. 3559-3564). IEEE.
- Kiesler, S. (2005). Fostering common ground in human-robot interaction. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication*, 2005. IEEE, 729–734.
- Kiesler, S., & Goetz, J. (2002). Mental models of robotic assistants. In *CHI'02 extended abstracts on Human Factors in Computing Systems* (pp. 576-577).
- Kiesler, S., & Hinds, P. (2004). Introduction to this special issue on human-robot interaction. *Human-Computer Interaction*, 19(1-2), 1–8. [https://doi.org/10.1207/s15327051hci1901&2\\_1](https://doi.org/10.1207/s15327051hci1901&2_1)
- Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2), 169–181.
- Kim, B., de Visser, E., & Phillips, E. (2022). Two uncanny valleys: Re-evaluating the uncanny valley across the full spectrum of real-world human-like robots. *Computers in human behavior*, 135, 107340.
- Kim, R. H., Moon, Y., Choi, J. J., & Kwak, S. S. (2014). The effect of robot appearance types on motivating donation. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction* (pp. 210-211).
- Kim, R. Y. (2024). Anthropomorphism and Human-Robot Interaction. *Communications of the ACM*, 67(2), 80-85.
- Kim, S. (2022). Working with robots: human resource development considerations in human–robot interaction. *Human Resource Development Review*, 21(1), 48-74.
- Kim, S. K., Kirchner, E. A., Schloßmüller, L., & Kirchner, F. (2020). Errors in human-robot interactions and their effects on robot learning. *Frontiers in Robotics and AI*, 7, 558531.
- Kirby, D. A. (2011). *Lab coats in Hollywood: Science, scientists, and cinema*. Cambridge, MA: MIT Press.
- Kirby, R., Forlizzi, J., & Simmons, R. (2010). Affective social robots. *Robotics and Autonomous Systems*, 58(3), 322-332.



- Kirchhoff, M. (2018). Predictive brains and embodied, enactive cognition: an introduction to the special issue. *Synthese*, 195(6), 2355-2366.
- Kirchhoff, M. D. (2015). Experiential fantasies, prediction, and enactive minds. *Journal of Consciousness Studies*, 22(3-4), 68-92.
- Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third wave view*. Routledge.
- Kirchhoff, M. D., & Kiverstein, J. (2020). Attuning to the world: The diachronic constitution of the extended conscious mind. *Frontiers in psychology*, 11, 1966.
- Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical explorations*, 21(2), 264-281.
- Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, 15(138), 20170792.
- Kiverstein, J. (2012). What is Heideggerian cognitive science?. In Kiverstein, J. & Wheeler, M. (Eds.), *Heidegger and Cognitive Science*, Houndmills, Basingstoke: Palgrave Macmillan, 1-61.
- Kiverstein, J. (2020). Free energy and the self: An ecological-enactive interpretation. *Topoi*, 39, 559-574. doi:10.1007/s11245-018-9561-5.
- Kiverstein, J., & Rietveld, E. (2021). Scaling-up skilled intentionality to linguistic thought. *Synthese*, 198(Suppl 1), 175-194.
- Kobayashi, Y., & Kuno, Y. (2010). People tracking using integrated sensors for human robot interaction. In *2010 IEEE International Conference on Industrial Technology* (pp. 1617-1622).
- Kodate N, Maeda Y, Hauray B, Tsujimura M, Chan WCH, Mannan H, Yu W, Dalgalarondo S, Cheung MC, Yumoto A, Suwa S, Donnelly S, Sakata N, O'Shea D, Obayashi K and Masuyama S (2022) Hopes and fears regarding care robots: Content analysis of newspapers in East Asia and Western Europe, 2001–2020. *Frontiers in Rehabilitation Sciences*, 3, 1019089.
- Kompatsiari, K., Ciardo, F., Tikhonoff, V., Metta, G., & Wykowska, A. (2021). It's in the eyes: The engaging role of eye contact in HRI. *International Journal of Social Robotics*, 13, 525-535.
- Kompatsiari, K., Tikhonoff, V., Ciardo, F., Metta, G., & Wykowska, A. (2017). The importance of mutual gaze in human-robot interaction. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9* (pp. 443-452). Springer International Publishing.
- Koopman, R. (2019). *The Uncanny Valley as a universal experience: a replication study using multilevel modelling* (Bachelor's thesis, University of Twente).
- Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., & Nehaniv, C. L. (2009). Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, 23(14), 1951-1996.
- Kotseruba, J., & Tsotsos, J. K. (2020). 40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications. *Artificial Intelligence Review* 53 (1): 17–94.

- Koyré, A. (1948) Du monde de l'«à-peu-près» à l'univers de la précision. *Critique*, 28 (tr. it. in *Dal mondo del pressappoco all'universo della precisione*, Torino, Einaudi, 90-111, 1967).
- Koyré, A. (1965). *Newtonian studies*. Cambridge, MA: Harvard University Press.
- Koyré, A. (1967) *Dal mondo del pressappoco all'universo della precisione*. Torino: Einaudi.
- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLOS ONE*, 3(7), e2597.
- Kruglanski, A. W., & Stroebe, W. (Eds.). (2012). *Handbook of the history of social psychology*. Psychology Press.
- Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(6), e1503.
- Kuipers, B., Feigenbaum, E. A., Hart, P. E., & Nilsson, N. J. (2017). Shakey: From Conception to History. *AI Magazine*, 38(1), 88-103. <https://doi.org/10.1609/aimag.v38i1.2716>
- Kupferberg A, Huber M, Helfer B, Lenz C, Knoll A, et al. (2012) Moving Just Like You: Motor Interference Depends on Similar Motility of Agent and Observer. *PLOS ONE* 7(6): e39637. <https://doi.org/10.1371/journal.pone.0039637>
- Kupferberg, A., Glasauer, S., Huber, M., Rickert, M., Knoll, A., & Brandt, T. (2011). Biological movement increases acceptance of humanoid robots as human partners in motor interaction. *AI & society*, 26, 339-345.
- Kwan, S. (2014). The impact of robot appearance type on social interaction with a robot and service evaluation of a robot. *Archives of design research*, 27(2), 81-93.
- Kyung-Joong, K. & Lipson, H. (2009). Towards a simple robotic theory of mind. In *Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems (PerMIS '09)*. Association for Computing Machinery, New York, NY, USA, 131–138. <https://doi.org.ezproxy.unibo.it/10.1145/1865909.1865937>
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago/London: University of Chicago Press.
- Langton, C. G. (1997). *Artificial Life: An Overview*. Cambridge, MA: MIT Press.
- Larghi, S., Datteri, E. (2024). Mentalistic Stances Towards AI Systems: Beyond the Intentional Stance. In Aldini, A. (eds) *Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops*. SEFM 2023. Lecture Notes in Computer Science, vol 14568. Springer, Cham.
- Lashley, K. S. (1951). *The problem of serial order in behavior*. Oxford: Bobbs-Merrill.
- Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.
- Lavazza, A. (2008). L'uomo a due dimensioni. *Il dualismo mente-corpo oggi*, Milano: Mondadori.
- Law, H. H. (1932). The Name Galatea in the Pygmalion Myth. *The Classical Journal*, 27(5), 337–342.

- Lay, S. (2015). *The Uncanny Valley Effect*. Open University.
- Le Cun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (1988, June). A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school* (Vol. 1, pp. 21-28).
- Lee, H. R., Šabanović, S., & Stolterman, E. (2016, March). How humanlike should a social robot be: a user-centered exploration. In *2016 AAAI Spring Symposium Series*.
- Lee, J. (2017). A survey of robot learning from demonstrations for human-robot collaboration. *arXiv preprint arXiv:1710.08789*.
- Lee, M., Lucas, G., & Gratch, J. (2021). Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator, and ultimatum games. *Journal on Multimodal User Interfaces*, 15(2), 201–214
- Lemaignan, S., Fink, J., Mondada, F., & Dillenbourg, P. (2015). You're doing it wrong! Studying unexpected behaviors in child-robot interaction. In *Proceedings of the International Conference on Social Robotics* (pp. 390–400).
- Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind.". *Psychological review*, 94(4), 412.
- Leslie, A. M. (1992). Pretense, autism, and the theory-of-mind module. *Current directions in psychological science*, 1(1), 18-21.
- Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In Hirschfeld, L. & Gelman, S. (Eds.) *Mapping the mind: Domain specificity in cognition and culture*, Cambridge, MA: Cambridge University Press, 119-148.
- Leslie, A. M. (2000). How to acquire a 'representational theory of mind'. *Metarepresentations: A multidisciplinary perspective*, 197-223.
- Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, 8(12), 528-533.
- Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11), 1940-1951.
- Levillain, F., & Zibetti, E. (2017). Behavioral objects: The rise of the evocative machines. *Journal of Human-Robot Interaction*, 6(1), 4-24.
- Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: the role of race categories. *Journal of Experimental Psychology: General*, 135(4), 501.
- Levin, D. T., Killingsworth, S. S., Saylor, M. M., Gordon, S. M., & Kawamura, K. (2013). Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human-Computer Interaction*, 28(2), 161–191
- Levin, D. T., Saylor, M. M., & Lynn, S. D. (2012). Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers. *International Journal of Human-Computer Studies*, 70(8), 527–534

- Lewkowicz, D. J., & Ghazanfar, A. A. (2012). The development of the uncanny valley in infants. *Developmental psychobiology*, 54(2), 124-132.
- Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34, No. 34).
- Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77, 23-37.
- Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Computers in Human Behavior*, 118, 106680.
- Lieto, A. (2021). *Cognitive design for artificial minds*. London: Routledge.
- Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6), 335-346.
- Lim, V., Rooksby, M. & Cross, E.S. Social Robots on a Global Stage: Establishing a Role for Culture During Human–Robot Interaction. *Int J of Soc Robotics* 13, 1307–1333 (2021). <https://doi.org/10.1007/s12369-020-00710-4>
- Lindblom, J. (2015). *Embodied social cognition* (Vol. 26). Berlin: Springer.
- Lindblom, J., & Ziemke, T. (2008). Interacting socially through embodied action. In *Enacting intersubjectivity* (pp. 49-63). Ios Press.
- Lindblom, J., Alenljung, B., Billing, E. (2020). Evaluating the User Experience of Human–Robot Interaction. In: Jost, C., et al. *Human-Robot Interaction*. Springer Series on Bio- and Neurosystems, vol 12. Springer, Cham. [https://doi.org/10.1007/978-3-030-42307-0\\_9](https://doi.org/10.1007/978-3-030-42307-0_9)
- Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., & Lederberg, J. (1980). *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. New York: McGraw-Hill.
- Lionel P. Robert Jr., Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You and Qiaoning Zhang (2020), "A Review of Personality in Human–Robot Interactions", *Foundations and Trends® in Information Systems*: Vol. 4: No. 2, pp 107-212. <http://dx.doi.org/10.1561/29000000018>
- Lobaccaro, L. (2022a). Ai confini del senso. La schizofrenia tra semiotica, psicopatologia e scienze cognitive. Tesi di Dottorato in *Philosophy, Science, Cognition and Semiotics*, Alma Mater Studiorum – Università di Bologna.
- Lobaccaro, L. (2022b). Default Mode Network, Schizophrenia, and Narrativity. Comments on Psychopathology of Language. *RETI SAPERI LINGUAGGI*, 9(2), 285-314 [10.12832/105800].
- Lobaccaro, L., & Bacaro, M., (2021). What is in the mirror? on mirror self-recognition, semiotics and material engagement. *Reti Saperi Linguaggi – Italian Journal of Cognitive Science*, 8, pp. 103-124.
- Locatelli, M., & Toniolo, F. (Eds) (2022). *Artificial lives: the humanoid robot in contemporary media culture*. Milano: Franco Angeli

- Logsdon, M. C., Popa, D., Mitchell, H., Chlebowy, D., Wijayasinghe, I., Saadatzi, M. N., & Das, S. (2019). Efficacy and acceptability of robots to improve physical tasks in hospital environments.
- Longo, G. O. (2009). Il test di Turing: storia e significato. *Mondo Digitale*, 1.
- Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological science*, 21(12), 1854-1862.
- Losano, M. (1990) *Storie di automi. Dalla Grecia classica alla Belle Époque*. Torino: Einaudi.
- Lu, L., Cai, R., & Gursoy, D. (2019). Developing and validating a service robot integration willingness scale. *International Journal of Hospitality Management*, 80, 36-51.
- Luisi, P. L. (2016). *The emergence of life: from chemical origins to synthetic biology*. Cambridge, MA: Cambridge University Press.
- Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection science*, 15(4), 151-190.
- Łupkowski, P., Rybka, M., Dziedzic, D., & Włodarczyk, W. (2017). Human-likeness assessment for the uncanny valley hypothesis. *Bio-Algorithms and Med-Systems*, 13(3), 125-131
- Lyons, J. (1995). *Approaches to intentionality*. Cambridge University Press.
- MacCormac, E. R. (1985). A Cognitive Theory of Metaphor. *Journal of Aesthetics and Art Criticism* 45 (4):418-420.
- MacDorman, K. F. (2005). Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it. In *CogSci-2005 workshop: toward social mechanisms of android science* (Vol. 3, pp. 106-118).
- MacDorman, K. F. (2006a). Introduction to the special issue on android science. *Connection Science*, 18(4), 313-317.
- MacDorman, K. F. (2006b). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science*.
- MacDorman, K. F. (2024). Does mind perception explain the uncanny valley? A meta-regression analysis and (de) humanization experiment. *Computers in Human Behavior: Artificial Humans*, 2(1), 100065.
- MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, 146, 190-205.
- MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, 16(2), 141-172.
- MacDorman, K. F., & Ishiguro, H. (2006a). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 7(3), 297-337.

- MacDorman, K. F., & Ishiguro, H. (2006b). Toward social mechanisms of android science. *Interaction Studies*, 7(2), 289-296.
- MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in human behavior*, 25(3), 695-710.
- Mahdi, H., Sami Alperen Akgun, Shahed Saleh, Kerstin Dautenhahn. (2022) A survey on the design and evolution of social robots — Past, present and future. *Robotics and Autonomous Systems*, Volume 156, 2022, 104193, ISSN 0921-8890, <https://doi.org/10.1016/j.robot.2022.104193>.
- Malafouris, L. (2013). *How things shape the mind*. Oxford University Press.
- Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious*. New York: Oxford University Press.
- Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In Chadee, D. (Ed.), *Theories in social psychology* (pp. 72-95). Wiley-Blackwell.
- Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology*, 33(2), 101-121.
- Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001). Introduction: The significance of intentionality. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 1-10). MIT Press.
- Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition*. MIT Press.
- Mara, M., Appel, M., & Gnambs, T. (2022). Human-like robots and the uncanny valley. *Zeitschrift für Psychologie*.
- Marchesi, S. (2021). *Individual and Cultural Differences in the Adoption of the Intentional Stance Towards Robots* (Order No. 29330224). Available from ProQuest One Academic; SciTech Premium Collection; Social Science Premium Collection. (2723854765). <https://www.proquest.com/dissertations-theses/individual-cultural-differences-adoption/docview/2723854765/se-2>
- Marchesi, S., & Spatola, N. (2024). Factors Influencing the Adoption of the Intentional Stance: Culture. In Wykowska, A. (eds) *Intentional Stance Towards Humanoid Robots*. Springer, Cham. [https://doi.org/10.1007/978-3-031-65483-1\\_6](https://doi.org/10.1007/978-3-031-65483-1_6)
- Marchesi, S., & Wykowska, A. (2024). Designing robots that are accepted in human social environments: Anthropomorphism, the intentional stance, cultural norms and values, and societal implications. In L. Fortunati & A. Edwards (Eds.), *The De Gruyter handbook of robots in society and culture* Berlin, Boston: De Gruyter, (pp. 63-84). De Gruyter. <https://doi.org/10.1515/9783110792270-004>
- Marchesi, S., Abubshait, A., Kompatsiari, K., Wu, Y., & Wykowska, A. (2023). Cultural differences in joint attention and engagement in mutual gaze with a robot face. *Scientific Reports*, 13(1), 11689.

- Marchesi, S., Bossi, F., Ghiglino, D., De Tommaso, D., & Wykowska, A. (2021c). I am looking for your mind: Pupil dilation predicts individual differences in sensitivity to hints of human-likeness in robot behavior. *Frontiers in Robotics and AI*, 8, 653537.
- Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots?. *Frontiers in psychology*, 10, 409635.
- Marchesi, S., Spatola, N., & Wykowska, A. (2021b). The mediating role of anthropomorphism in adopting the intentional stance towards humanoid robots. PsyArXiv; 2021. DOI: 10.31234/osf.io/zdr9v
- Marchesi, S., Spatola, N., Perez-Osorio, J., & Wykowska, A. (2021a). Human vs humanoid. A behavioral investigation of the individual tendency to adopt the intentional stance. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 332-340).
- Marchesi, S., Spatola, N., Wykowska, A. (2024). Delineation and Operationalisation of the Concept of the Intentional Stance. In: Wykowska, A. (Ed.) *Intentional Stance Towards Humanoid Robots*. Springer, Cham. [https://doi.org/10.1007/978-3-031-65483-1\\_2](https://doi.org/10.1007/978-3-031-65483-1_2)
- Marchis, V. (2014) Uomini, macchine, automi. In Eco U., Corsi, P. (Eds.) *Il Settecento - Scienze e tecniche*. Milano: EM Publishers (e-book).
- Margoni, F., Surian, L., & Baillargeon, R. (2024). The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3), 716.
- Marr, D. (1982) *Vision. A computational investigation into the human representation and processing of visual information*. MIT Press.
- Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1156), 301-328.
- Marraffa, M., & Paternoster, A. (2012). *Persone, menti, cervelli. Storia, metodi e modelli delle scienze della mente*. Milano: Mondadori.
- Martin, D. U., Perry, C., MacIntyre, M. I., Varcoe, L., Pedell, S., & Kaufman, J. (2020). Investigating the nature of children's altruism using a social humanoid robot. *Computers in Human Behavior*, 104, 106149.
- Martini, M. C., Gonzalez, C. A., & Wiese, E. (2016). Seeing minds in others—Can agents with robotic appearance have human-like preferences?. *PloS one*, 11(1), e0146310.
- Maruyama, N., Masumori, A., & Ikegami, T. (2021, July). Studying embodied cognition in the android Alter. In *ALIFE 2021: The 2021 conference on artificial life*. MIT Press.
- Masuda T., Nisbett R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81, 922–934.
- Matarić, M. J. (1998). Behavior-based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior. *Trends in cognitive sciences*, 2(3), 82-86.
- Matarić, M. J. (2007). *The robotics primer*. Cambridge, MA: MIT press.

- Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, 146, 22-32.
- Mathur, M. B., Reichling, D. B., Lunardini, F., Geminiani, A., Antonietti, A., Ruijten, P. A., ... & Aczel, B. (2020). Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley. *Computers in Human Behavior*, 103, 21-30.
- Matsuda, Y. T., Okamoto, Y., Ida, M., Okanoya, K., & Myowa-Yamakoshi, M. (2012). Infants prefer the faces of strangers or mothers to morphed faces: an uncanny valley between social novelty and familiarity. *Biology letters*, 8(5), 725-728.
- Maturana, H. (1970). *Biology of cognition*. Urbana: Biological Computer Laboratory, Department of Electrical Engineering, University of Illinois (tr. it. in Maturana, H. & Varela, F. *Autopoiesi e cognizione. La realizzazione del vivente*, Venezia, Marsilio, 1985, 47-110).
- Maturana, H. & Varela, F. (1972) *De Maquinas y seres vivos. Una teoria sobre la organización biológica*. Editorial Universitaria (tr. it. *Macchine ed esseri viventi. L'autopoiesi e l'organizzazione biologica*, Udine, Astrolabio, 1992).
- Maturana, H. & Varela, F. (1973) *Autopoiesi. L'organizzazione del vivente*, ora in Maturana, H. & Varela, F. *Autopoiesi e cognizione. La realizzazione del vivente*, Venezia, Marsilio, 1985, 111-198.
- Maturana, H. & Varela, F. (1980) *Autopoiesis and Cognition. The realization of the living*. Dordrecht: Reider Publishing (tr. it. *Autopoiesi e cognizione. La realizzazione del vivente*, Venezia, Marsilio, 1985).
- Maurice, M. E. Huber, N. Hogan and D. Sternad, "Velocity-Curvature Patterns Limit Human-Robot Physical Interaction," in *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 249-256, Jan. 2018, doi: 10.1109/LRA.2017.2737048.
- Mayor, A. (2018). *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton: Princeton University Press.
- McCarthy, J & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4: 463–502
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. *AI Magazine*, 27(4), 12.
- McCulloch, W. S. (1965). *Embodiments of Mind*. Cambridge (MA): MIT Press.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115-133.
- McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4), 1-11.
- McGann, M., & De Jaegher, H. (2009). Self-other contingencies: Enacting social perception. *Phenomenology and the cognitive sciences*, 8, 417-437.



- Meinong, A. (1960). *On Objects of Higher Order*. In R. Chisholm (Ed.), *The Theory of Knowledge* (pp. 1-22). New York: Harper & Row.
- Meister, M. (2014). When is a robot really social? An outline of the robot sociologicus. *Science, Technology & Innovation Studies*, 10(1), 107-134.
- Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of mind. *The Wiley-Blackwell handbook of childhood cognitive development*, 2, 49-75.
- Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. (2010). “Social” robots are psychological agents for infants: A test of gaze following. *Neural networks*, 23(8-9), 966-972.
- Meltzoff, A., & Moore, M. (1997) Explaining facial imitation: a theoretical model. *Early Development and Parenting*, 6: 179–192.
- Merleau-Ponty, M. (1942). *La structure du comportement*. Paris: Presses Universitaires de France (tr. it. *La struttura del comportamento*, Mimesis, Milano, 2019).
- Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Paris: Gallimard (tr. it. *Fenomenologia della percezione*, Bompiani, Milano, 2003)
- Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., ... & Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, 23(8-9), 1125-1134.
- Mettrie, J. O. de La (1747). *L’Homme machine* (tr. it. *L’uomo macchina e altri scritti*. Milano, Feltrinelli, 1955).
- Milkowski, M. (2013). *Explaining the computational mind*. MIT Press.
- Miller, M. (2013) *Galatea: A Short Story*. New York: Ecco.
- Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*.
- Minh Trieu, N., & Truong Thinh, N. (2023). A comprehensive review: Interaction of appearance and behavior, artificial skin, and humanoid robot. *Journal of Robotics*, 2023(1), 5589845.
- Mininni. G. (2000). Intenzionalità. In *Enciclopedia Italiana Treccani - VI Appendice*: [https://www.treccani.it/enciclopedia/intenzionalita\\_\(Enciclopedia-Italiana\)/](https://www.treccani.it/enciclopedia/intenzionalita_(Enciclopedia-Italiana)/)
- Minsky, M. (1961) Steps toward Artificial Intelligence. *Proceeding of the Institute of Radio Engineers*, 49: 8-30; pubblicato originariamente come MIT Technical Report (1956).
- Minsky, M. (1974) A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306, June, 1974. Ora in versione ridotta in Haugeland, J. (Ed.), *Mind Design*, MIT Press, 1981.
- Minsky, M. (1986) *The Society of Mind*. New York: Simon & Schuster (tr. it. *La società della mente*, Milano, Adelphi, 1989).
- Minsky, M. (Ed.) (1985) *Robotics: The First Authoritative Report from the Ultimate High-Tech Frontier*. New York Doubleday. (tr. it. Sosio, L., *La robotica. Il primo autorevole rapporto delle frontiere dell’alta tecnologia*. Milano: Longanesi, 1987).

- Minsky, M. & Papert, S. (1969) *Perceptrons*. Cambridge, MA: MIT Press.
- Minsky, M. L. (1982). Why People Think Computers Can't. *AI Magazine*, 3(4), 3.
- Miraglia, L., Peretti, G., Manzi, F., Di Dio, C., Massaro, D., & Marchetti, A. (2023). Development and validation of the Attribution of Mental States Questionnaire (AMS-Q): A reference tool for assessing anthropomorphism. *Frontiers in psychology*, 14, 999921.
- Mishra, N., Ramanathan, M., Tulsulkar, G., & Thalmann, N. M. (2022). Uncanny valley for interactive social agents: An experimental study. *Virtual Reality & Intelligent Hardware*, 4(5), 393-405.
- Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines*, 19, 345-359.
- Mitchell, J. (2005). Theory as an Object. *October*, 113, 27–38. <http://www.jstor.org/stable/3397652>
- Mitchell, M. (2002). Imitation as a perceptual process. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in Animals and Artifacts* (pp. 441–469). MIT Press.
- Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, 2(1), 10-12.
- Miyahara, K. (2011). Neo-pragmatic intentionality and enactive perception: A compromise between extended and enactive minds. *Phenomenology and the Cognitive Sciences*, 10(4), 499-519.
- Mohammad, Y., & Nishida, T. (2015). *Data mining for social robotics*. Cham, Switzerland: Springer.
- Moon, A., Troniak, D. M., Gleeson, B., Pan, M. K. X. J., Zheng, M., Blumer, B. A., MacLean, K., & Croft, E. A. (2014). Meet me where I'm gazing: How shared attention gaze affects human-robot handover timing. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI '14) (pp. 334–341). Association for Computing Machinery. <https://doi.org/10.1145/2559636.2559656>
- Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific reports*, 2(1), 864.
- Moran, D. (2013). Intentionality: Some Lessons from the History of the Problem from Brentano to the Present. *International Journal of Philosophical Studies*, 21(3), 317-358.
- Moravec, H. P. (1980). *Obstacle avoidance and navigation in the real world by a seeing robot rover*. Stanford University.
- Moravec, H. P. (1988). *Mind children: The future of robot and human intelligence*. Cambridge, MA: Harvard University Press.
- Moravec, H. P. (1999). *Robot: mere machine to transcendent mind*. Cambridge, MA: Oxford University Press.
- Morganti, F., Carassa, A., & Riva, G. (Eds.). (2008). *Enacting intersubjectivity: A cognitive and social perspective on the study of interactions* (Vol. 10). Ios Press.

- Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, 7:33–35.
- Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2), 98-100.
- Morin, A. (2006). Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and Cognition*, 15(2), 358-371. <https://doi.org/10.1016/j.concog.2005.09.006>
- Morin, E. (1977). *Le Méthode. 1. La Nature de la Nature*. Paris: Éditions du Seuil (tr. it. *Il Metodo. 1. La natura della natura*, Milano, Cortina, 2001).
- Morin, E. (2011) *La sfida della complessità*. Firenze: Le Lettere.
- Morowitz, H. (1980) Rediscovering the Mind. *Psychology Today*. In Hofstadter, D. & Dennett, D. (Eds.) (1981) *The Mind's I. Fantasies and Reflections on Self and Soul*. New York: Basic Books (tr. it. *L'io della mente*, Milano, Adelphi, 1985).
- Mosavi, A., & Varkonyi, A. (2017). Learning in robotics. *International Journal of Computer Applications*, 157(1), 8-11.
- Moschovakis, Y. N. (2001). What Is an Algorithm?. In Engquist, B., Schmid, W. (eds) *Mathematics Unlimited — 2001 and Beyond*. Heildelberg: Springer, Berlin, 919-936.
- Mou, W., Ruocco, M., Zanatto, D., & Cangelosi, A. (2020, August). When would you trust a robot? a study on trust and theory of mind in human-robot interactions. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 956-962). IEEE.
- Mumford, L. (1961). History: neglected clue to technological change. *Technology and Culture*, 2(3), 230-236.
- Murdoch, I. (1999). Metaphysics and ethics. In P. Conradi (ed.), *Existentialists and Mystics*, 59–75, New York:Penguin.
- Murray, L., & Trevarthen, C. (1986). The infant's role in mother–infant communications. *Journal of child language*, 13(1), 15-29.
- Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Proceedings of the 2009 4th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 69–76).
- Nannini, S. (2023). The mind-body problem between philosophy and the cognitive sciences. *Rivista internazionale di Filosofia e Psicologia*, 14(1-2), 118-134.
- Nasir, J., Oppliger, P., Bruno, B., & Dillenbourg, P. (2022, August). Questioning wizard of Oz: Effects of revealing the wizard behind the robot. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1385-1392). IEEE.
- Navarretta, C. (2014). Predicting emotions in facial expressions from the annotations in naturally occurring first encounters. *Knowledge-Based Systems*, 71, 34-40.

- Navarretta, C. (2016). Mirroring facial expressions and emotions in dyadic conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 469-474).
- Nehaniv, C. L., & Dautenhahn, K. (2002). The correspondence problem. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in Animals and Artifacts* (pp. 41–61). MIT Press.
- Neisser, U. (1967) *Cognitive Psychology*. New York: Appleton-Century-Crofts.
- Neumann, J. von (1958) *The Computer and the Brain*. New Haven, CT: Yale University Press.
- Newel, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113-126.
- Newell, A. & Simon, H. (1956) The Logic Theory Machine: A Complex Information-Processing System. *IRE Transactions on Information Theory*, IT-2/3: 61-79.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.
- Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological review*, 65(3), 151.
- Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. In *IFIP congress*, Vol. 256, 64.
- Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press.
- Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge (MA): Cambridge University Press.
- Nilsson, N. J. (Ed.). (1984). *Shakey the robot*. Menlo Park, California: SRI International.
- Nisbett, R. (2003) *The Geography of thought*. Free Press.
- Nishio, S., Ishiguro, H., & Hagita, N. (2007). Geminoid: Teleoperated android of an existing person. *Humanoid robots: New developments*, 14(343-352), 10-1109.
- Nisimura, R., Uchida, T., Lee, A., Saruwatari, H., Shikano, K., & Matsumoto, Y. (2002). ASKA: receptionist robot with speech dialogue system. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Lausanne, Switzerland, 2002, pp. 1314-1319 vol.2, doi: 10.1109/IRDS.2002.1043936.
- Noceti, N., Sciutti, A., & Sandini, G. (2015). Cognition helps vision: Recognizing biological motion using invariant dynamic cues. In *Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part II* 18 (pp. 676-686). Springer International Publishing.
- Noë, A. (2004). *Action in perception*. Cambridge (MA): MIT Press.
- Nolfi, S., Bongard, J., Husbands, P., & Floreano, D. (2016). Evolutionary robotics. In Siciliano, B. & Khatib, O. (Eds.) (2016) *Springer Handbook of Robotics*, pp. 2035-2068: Springer International Publishing.

- Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai & Society*, 20, 138-150.
- Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.
- Nyholm, S. (2023). Artificial intelligence, humanoid robots, and old and new control problems. In *Social Robots in Social Institutions* (pp. 3-12). IOS Press.
- Oatley, K., & Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *British journal of social psychology*, 24(2), 115-124.
- Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, 70(13-15), 2194-2203.
- Obrenovic, B., Gu, X., Wang, G. et al. Generative AI and human–robot interaction: implications and future agenda for business, society and ethics. *AI & Soc* (2024). <https://doi.org/10.1007/s00146-024-01889-0>
- Odifreddi, P. (2000) *La matematica del Novecento. Dagli insiemi alla complessità*. Einaudi.
- Ogden, B., & Dautenhahn, K. (2000). Robotic etiquette: Structured interaction in humans and robots. In *Procs sirs 2000, 8th symp on intelligent robotic systems*. University of Reading.
- Onnasch, L., & Roesler, E. (2021). A taxonomy to structure and analyze human–robot interaction. *International Journal of Social Robotics*, 13(4), 833-849.
- Ono, T., Imai, M., & Nakatsu, R. (2000). Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. *Advanced Robotics*, 14(4), 311-326.
- Osawa, H., Mukai, J., & Imai, M. (2007). Anthropomorphization framework for human-object communication. *Journal of Advanced Computational Intelligence and Intelligence Informatics*, 11(8), 1007-1014.
- Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2017). Believing androids: fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, 12(5), 582–593.
- Paetzel, M., & Castellano, G. (2019, September). Let me get to know you better: can interactions help to overcome uncanny feelings?. In *Proceedings of the 7th International Conference on Human-agent Interaction* (pp. 59-67).
- Paetzel, M., Perugia, G., & Castellano, G. (2020, March). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 73-82).
- Paolucci, C. (2010). *Strutturalismo e interpretazione*. Milano: Bompiani.
- Paolucci, C. (2020). A Radical Enactivist Approach to Social Cognition. In Pennisi, A., Falzone, A. (eds) *The Extended Theory of Cognitive Creativity. Perspectives in Pragmatics, Philosophy & Psychology*, vol 23. Springer, Cham.

- Paolucci, C. (2021). *Cognitive semiotics*. Springer International Publishing.
- Papadopoulos I, Koulouglioti C. The Influence of Culture on Attitudes Towards Humanoid and Animal-like Robots: An Integrative Review. *J Nurs Scholarsh*. 2018 Nov;50(6):653-665. doi: 10.1111/jnu.12422. Epub 2018 Sep 21. PMID: 30242796.
- Parmiggiani, A., Maggiali, M., Natale, L., Nori, F., Schmitz, A., Tsagarakis, N., ... & Metta, G. (2012). The design of the iCub humanoid robot. *International journal of humanoid robotics*, 9(04), 1250027
- Parviainen, J., Turja, T., & Van Aerschot, L. (2018, November). Robots and human touch in care: desirable and non-desirable robot assistance. In *International Conference on Social Robotics* (pp. 533-540). Cham: Springer International Publishing.
- Pekçetin, T. N., Acarturk, C., & Urgan, B. A. (2024, March). Investigating Mind Perception in HRI through Real-Time Implicit and Explicit Measurements. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 139-141).
- Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122, 106855.
- Peled-Avron, L., Perry, A., & Shamay-Tsoory, S. G. (2016). The effect of oxytocin on the anthropomorphism of touch. *Psychoneuroendocrinology*, 66, 159-165.
- Pelikan, H. & Broth, M. (2016). Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4921–4932. <https://doi.org/10.1145/2858036.2858478>
- Pereira, A., Oertel, C., Feroselle, L., Mendelson, J., & Gustafson, J. (2020, March). Effects of different interaction contexts when evaluating gaze models in HRI. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 131-139).
- Perez-Osorio, J., & Wykowska, A. (2019). Adopting the intentional stance towards humanoid robots. *Wording robotics: Discourses and representations on robotics*, 119-136.
- Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3), 369-395.
- Peters, J. C., Goebel, R., & Goffaux, V. (2018). From coarse to fine: Interactive feature processing precedes local feature analysis in human face perception. *Biological psychology*, 138, 1-10.
- Petrovych, V., Thellman, S., & Ziemke, T. (2018). Human interpretation of goal-directed autonomous car behavior. In *The 40th annual cognitive science society meeting, Madison, Wisconsin, USA, July 25-28* (pp. 2235-2240). Cognitive Science Society.
- Pfeifer, R. (1995). Cognition—perspectives from autonomous agents. *Robotics and Autonomous Systems*, 15(1-2), 47-70.
- Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. Cambridge, MA: MIT press.



- Pfeifer, R., & Scheier, C. (2001). *Understanding intelligence*. Cambridge, MA: MIT press.
- Pfeiffer, U. J., Timmermans, B., Bente, G., Vogeley, K., & Schilbach, L. (2011). A non-verbal turing test: differentiating mind from machine in gaze-based social interaction. *PloS one*, 6(11), e27591.
- Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In HRI '18: 2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5–8, 2018, Chicago, IL, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3171221.3171268>
- Piccolino, M. e Bresaola, M. (2003), *Rane, torpedini e scintille, Galvani, Volta e l'elettricità animale*. Torino: Bollati Boringhieri.
- Pino, O., Palestra, G., Trevino, R., & De Carolis, B. (2020). The humanoid robot NAO as trainer in a memory program for elderly people with mild cognitive impairment. *International Journal of Social Robotics*, 12, 21-33.
- Piwek, L., McKay, L. S., & Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition*, 130(3), 271-277.
- Poliakoff, E., Beach, N., Best, R., Howard, T., & Gowen, E. (2013). Can looking at a hand make your skin crawl? Peering into the uncanny valley for hands. *Perception*, 42(9), 998-1000.
- Pollack, J. B., & Lipson, H. (2000). The GOLEM project: Evolving hardware bodies and brains. In *Proceedings. The Second NASA/DoD Workshop on Evolvable Hardware* (pp. 37-42).
- Pollick, F. E. (2009). In search of the uncanny valley. In *User Centric Media: First International Conference, UCMedia 2009, Venice, Italy, December 9-11, 2009, Revised Selected Papers 1* (pp. 69-78). Springer Berlin Heidelberg.
- Polson, K. & Scott, J. (2018) *AIQ: How People and Machines Are Smarter Together*. (tr. it. *Numeri intelligenti*, Milano, UTET, 2018).
- Povinelli, D. J. (1994). A theory of mind is in the head, not the heart. *Behavioral and Brain Sciences*, 17(3), 573-574.
- Prasad, A. (2014) *Storia naturale del concepimento*. Torino: Bollati Boringhieri.
- Premack D, Woodruff G. Does the chimpanzee have a theory of mind?. *Behav Brain Sci*. 1978;1: 515–526.
- Prescott, T. J. (2017). Robots are not just tools. *Connection Science*, 29(2), 142-149.
- Prigogine, I., Nicolis, G. (1985). Self-Organisation in Nonequilibrium Systems: Towards A Dynamics of Complexity. In Hazewinkel, M., Jurkovich, R., Paelinck, J.H.P. (eds) *Bifurcation Analysis*. Springer, Dordrecht, 3-12.
- Prince, C. G., & Gogate, L. J. (2007). Epigenetic robotics: behavioral treatments and potential new models for developmental pediatrics. *Pediatric research*, 61(4), 383-385.

- Putnam, H. (1960). Minds and Machines. In Hook, S. (ed.), *Dimensions of Mind: A Symposium*. New York: New York University Press, 138-164.
- Putnam, H. (1967). Psychological Predicates. In W.H. Capitan & D.D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37–48). University of Pittsburgh Press.
- Putnam, H. (1981). *Reasons, Truth, and History*. Cambridge (MA): Cambridge University Press.
- Pylyshyn, Z. W. (1978). When is the attribution of beliefs justified?. *The Behavioral and Brain Sciences*, 1(4): 592-593.
- Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain sciences*, 3(1), 111-132.
- Pylyshyn, Z. W. (ed.) (1987). *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Ablex.
- Pyszczynski, T., Greenberg, J., & Solomon, S. (1999). A dual-process model of defense against conscious and unconscious death-related thoughts: an extension of terror management theory. *Psychological review*, 106(4), 835.
- Qingxue, L. (2003). Understanding different cultural patterns or orientations between East and West. *Investigationes Linguisticae*, 9, 22-30.
- Quine, W. V. O. (1960). *Word and object*. MIT Press.
- Quintili, P. (Eds.) (2024) *Mind-Body Problem. Mente, corpo, emozioni e passioni*. Milano: Mimesis.
- Ramey, C. H. (2006). An inventory of reported characteristics for home computers, robots, and human beings: Applications for android science and the uncanny valley. In *Proceedings of the ICCS/CogSci-2006 Long Symposium 'Toward Social Mechanisms of Android Science', Vancouver, Canada*.
- Ratcliffe, M. (2017). Empathy without simulation. In Summa, M., Fuch, T. and Vanzago L. (eds.) *Imagination and Social Perspectives: Approaches from Phenomenology and Psychopathology*. London: Routledge, pp. 274-306.
- Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B*, 374(1771), 20180033.
- Ravichandar, H., Polydoros, A. S., Chernova, S., & Billard, A. (2020). Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3(1), 297-330.
- Rea, F., Vignolo, A., Sciutti, A., & Noceti, N. (2019). Human motion understanding for selecting action timing in collaborative human-robot interaction. *Frontiers in Robotics and AI*, 6, 58.
- Recio, D. L., Segura, L. M., Segura, E. M., & Waern, A. (2013). The NAO models for the elderly. In *2013 8th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 187-188). IEEE.
- Reddy, V. (2008) *How infants know minds*. Harvard University Press.



- Reddy, V., & Morris, P. (2004). Participants don't need theories: Knowing minds in engagement. *Theory & Psychology*, 14(5), 647-665.
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10), 19-36.
- Reichardt, J. (1978). *Robots: Fact, fiction, and prediction*. Viking Penguin.
- Riechelmann, E., Raettig, T., Böckler, A., & Huestegge, L. (2021). Gaze interaction: Anticipation-based control of the gaze of others. *Psychological research*, 85, 302-321.
- Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1), 119-136.
- Rietveld, E. (2012). Bodily intentionality and social affordances in context. In Paglieri, F. (Ed.), *Consciousness in interaction: The role of the natural and social context in shaping consciousness*, Amsterdam: John Benjamins, 207-226.
- Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological psychology*, 26(4), 325-352.
- Rizzolatti, G. (2005). The mirror neuron system and its function in humans. *Anatomy and embryology*, 210(5), 419-421.
- Rizzolatti, G., & Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Current opinion in neurobiology*, 18(2), 179-184.
- Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (1999). Resonance behaviors and mirror neurons. *Archives italiennes de biologie*, 137(2), 85-100.
- Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (2002). From mirror neurons to imitation: Facts and speculations. In A. N. Meltzoff & W. Prinz (Eds.), *The Imitative Mind. Development, Evolution, and Brain Bases* (pp. 247–265). Cambridge: Cambridge University Press.
- Robertson, J. (2018). *Robo sapiens japonicus: robots, gender, family, and the Japanese nation*. Univ of California Press.
- Robinson, H. (2023) Dualism. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2023/entries/dualism/>
- Rochat, P., Neisser, U., & Marian, V. (1998). Are young infants sensitive to interpersonal contingency?. *Infant behavior and development*, 21(2), 355-366.
- Roesler E, Manzey D, Onnasch L (2021) A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Sci Robot* 6(58), p eabj5425. <https://doi.org/10.1126/scirobotics.abj5425>
- Romeo, M, McKenna, PE, Robb, D, Rajendran, G, Nasset, B, Cangelosi, A & Hastie, H. (2022). Exploring Theory of Mind for Human-Robot Collaboration. In *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 461-468, 31st IEEE International Conference on Robot and Human Interactive Communication 2022, Napoli, Italy, 29/08/22. <https://doi.org/10.1109/RO-MAN53752.2022.9900550>

- Rosander, K., & von Hofsten, C. (2011). Predictive gaze shifts elicited during observed and performed actions in 10-month-old infants and adults. *Neuropsychologia*, 49, 2911–2917.
- Rosati, A. D., Knowles, E. D., Kalish, C. W., Gopnik, A., Ames, D. R., & Morris, M. W. (2001). What theory of mind can teach social psychology: traits as intentional terms. *Intentions and intentionality: Foundations of social cognition*, 287-303.
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3), 192.
- Rosch, E., & Lloyd, B. (1978). *Cognition and Categorization*. John Wiles and Sons.
- Roselli, C., Ciardo, F. & Wykowska, A. (2022). Social inclusion of robots depends on the way a robot is presented to observers. *Paladyn, Journal of Behavioral Robotics*, 13(1), 56-66. <https://doi.org/10.1515/pjbr-2022-0003>
- Roselli, C., Marchesi, S., De Tommaso, D., & Wykowska, A. (2023). The role of prior exposure in the likelihood of adopting the intentional stance toward a humanoid robot. *Paladyn, Journal of Behavioral Robotics*, 14(1), 20220103.
- Rosenblueth, A., & Wiener, N. (1945). The role of models in science. *Philosophy of science*, 12(4), 316-321 (tr. it. in Somenzi, V. & Cordeschi, R. *La filosofia degli automi. Origini dell'intelligenza artificiale*, Torino: Bollati, 1941, 2° edizione).
- Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of science*, 10(1), 18-24.
- Rosenschein, S. J., & Kaelbling, L. P. (1995). A situated view of representation and control. *Artificial intelligence*, 73(1-2), 149-173.
- Rosenthal-Von Der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. *Computers in Human Behavior*, 36, 422-439.
- Rosenthal-von der Pütten, A. M., Bock, N., Hoffmann, L., Krämer, N. C., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1), 17-34.
- Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., ... & Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior*, 33, 201-212.
- Rosenthal-von der Pütten, A., & Weiss, A. (2015). The uncanny valley phenomenon: Does it affect all of us. *Interaction Studies*, 16(2), 206-214.
- Rossi, C. & Russo, F. (2009). *Ancient Engineers' Inventions: Precursors of the Present*. Cham: Springer.
- Rossi, P. (1962) *I filosofi e le macchine: 1400-1700*. 3° edizione, 2009. Milano: Feltrinelli.
- Rowlands, M. (2009). The mind–body problem. In Banks, W. P. (Ed.) *Encyclopedia of consciousness*, Miami: Academic Press, 43-55.

- Rozin, P., & Fallon, A. E. (1987). A perspective on disgust. *Psychological review*, 94(1), 23.
- Ruijten, P. A., Haans, A., Ham, J., & Midden, C. J. (2019). Perceived human-likeness of social robots: testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics*, 11, 477-494.
- Russell, S. & Norvig, P. (2010). *Artificial intelligence: a modern approach*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.
- Ryle, G. (1949). *The concept of mind*. London: Hutchinson.
- Sachs, C. B. (2014). Discursive and somatic intentionality: Merleau-Ponty contra “McDowell or Sellars.” *International Journal of Philosophical Studies*, 22(2), 199–227.
- Saerbeck, M., & Bartneck, C. (2010). Perception of affect elicited by robot motion. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 53-60). IEEE.
- Saffari, E., Hosseini, S. R., Taheri, A., & Meghdari, A. (2021). Does cinema form the future of robotics?: a survey on fictional robots in sci-fi movies. *SN Applied Sciences*, 3(6), 655.
- Salemi, B., Reis, J., Saifhashemi, A., & Nikgozar, F. (2005, August). MILO: Personal robot platform. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 4089-4094). IEEE.
- Sandini, G., Sciutti, A., & Vernon, D. (2021). Cognitive robotics. In *Encyclopedia of Robotics* (pp. 1-7). Springer-Verlag.
- Sansoni, S., Wodehouse, A., McFadyen, A. K., & Buis, A. (2015). The aesthetic appeal of prosthetic limbs and the uncanny valley: The role of personal characteristics in attraction. *International Journal of Design*, 9(1), 67-81.
- Saran, A., Majumdar, S., Short, E. S., Thomaz, A., & Niekum, S. (2018a). Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 8615-8621). IEEE.
- Saran, A., Majumdar, S., Thomaz, A., & Niekum, S. (2018b). Real-time human gaze following for human-robot interaction. In *Proceedings of the International Conference on Human Robot Interaction*.
- Sarrica, M., Brondi, S., Fortunati, L. (2019). How many facets does a “social robot” have? A review of scientific and popular definitions online. *Information Technology & People*, (), ITP-04-2018-0203–. doi:10.1108/ITP-04-2018-0203
- Sartori, L., Becchio, C., & Castiello, U. (2011). Cues to intention: the role of movement information. *Cognition* 119, 242–252. doi:10.1016/j.
- Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, 7(4), 413-422.
- Scassellati, B. (1996). Mechanisms of shared attention for a humanoid robot. In *Embodied cognition and action: papers from the 1996 AAAI fall symposium* (Vol. 4, No. 9, p. 21).

- Scassellati, B. (1998). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *International Workshop on Computation for Metaphors, Analogy, and Agents* (pp. 176-195). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Scassellati, B. (2001). *Foundations for a Theory of Mind for a Humanoid Robot* (Doctoral dissertation, Massachusetts Institute of Technology).
- Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12, 13-24.
- Scassellati, B. (2003). Investigating models of social development using a humanoid robot. In *Proceedings of the International Joint Conference on Neural Networks, 2003*. (Vol. 4, pp. 2704-2709). IEEE.
- Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends Cogn. Sci.*, 3:233–42
- Schaerer, E., Kelley, R., & Nicolescu, M. (2009). Robots as animals: A framework for liability and responsibility in human-robot interactions. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, Toyama, Japan, 2009, pp. 72-77, doi: 10.1109/ROMAN.2009.5326244.
- Scheele, D., Schwering, C., Elison, J. T., Spunt, R., Maier, W., & Hurlemann, R. (2015). A human tendency to anthropomorphize is enhanced by oxytocin. *European Neuropsychopharmacology*, 25(10), 1817-1823.
- Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots—open questions and methodological challenges. *Frontiers in Robotics and AI*, 5, 139.
- Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the “default system” of the brain. *Consciousness and cognition*, 17(2), 457-467.
- Schilhab, T. S. (2002). Anthropomorphism and mental state attribution. *Animal behaviour*, 63(5), 1021-1026.
- Schlesinger, M. (2009). The robot as a new frontier for connectionism and dynamic systems theory. In John Spencer (Ed.) *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*, New York, Oxford Academic, 182-199.
- Schlicht, T. (2023). Predictive Processing Theories. In: *Philosophy of Social Cognition*. Palgrave Philosophy Today. Palgrave Macmillan, Cham. [https://doi.org/10.1007/978-3-031-14491-2\\_7](https://doi.org/10.1007/978-3-031-14491-2_7)
- Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). Retrieved from <https://plato.stanford.edu/archives/win2019/entries/agency/>
- Schmidtler, J., Knott, V., Hölzel, C., & Bengler, K. (2015). Human centered assistance applications for the working environment of the future. *Occupational Ergonomics*, 12(3), 83-95.
- Schofield, D., & LeRoy, N. C. (2018). Representing robots: the appearance of artificial humans in cinematic media. *Journal of Arts and Humanities*, 7(5), 12-28.)
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in cognitive sciences*, 4(8), 299-309.

- Scholtz, J. (2003) Theory and evaluation of human robot interactions. In *Proceedings of the 36th annual Hawaii international conference on system science*.
- Schrödinger, E. (1944) *What is life? The physical aspect of the living cell*. Cambridge, MA: Cambridge University Press (tr. it. *Cos'è la vita?*, Milano, Adelphi, 1995).
- Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, 145(11), 1427.
- Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological science*, 28(12), 1745-1762.
- Schwind, V. (2015). Historical, Cultural, and Aesthetic Aspects of the Uncanny Valley. *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*, 81-107.
- Schwind, V., Wolf, K., & Henze, N. (2018). Avoiding the uncanny valley in virtual character design. *interactions*, 25(5), 45-49.
- Sciutti, A., & Sandini, G. (2017). Interacting with robots to investigate the bases of social interaction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12), 2295-2304.
- Sciutti, A., Ansuini, C., Becchio, C., & Sandini, G. (2015). Investigating the ability to read others' intentions using humanoid robots. *Frontiers in psychology*, 6, 148051.
- Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., & Sandini, G. (2012b). Anticipatory gaze in human-robot interactions. In *Gaze in HRI from modeling to communication*, Workshop at the 7th ACM/IEEE international conference on human-robot interaction, Boston, Massachusetts, USA
- Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., & Sandini, G. (2013). Robots can be perceived as goal-oriented agents. *Interaction Studies*, 14(3), 329-350.
- Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., Pozzo, T., & Sandini, G. (2012a). Measuring human-robot interaction through motor resonance. *International Journal of Social Robotics*, 4(3), 223–234. <https://doi.org/10.1007/s12369-012-0143-1>
- Sciutti, A., Patanè, L., Nori, F., & Sandini, G. (2014). Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development*, 6, 80–92.
- Sciutti, F., Mara, M., Tagliasco, V., & Sandini, G. (2018). "Humanizing human-robot interaction: On the importance of mutual understanding." *IEEE Technology and Society Magazine*, 37(1), 22-29.
- Sciutti, F., Sandini, G., & Wykowska, A. (2021). "Machines that experience, understand us, and are self-aware: A roadmap toward human–robot symbiosis." *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), 503-512.
- Scott, B. (2004). Second-order cybernetics: an historical introduction. *Kybernetes*, 33(9/10), 1365-1378.

- Searle, J. R. (1980b). Intrinsic intentionality. *Behavioral and Brain Sciences*, 3(3), 450-457.
- Searle, J. R. (1980a). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417-424.
- Searle, J. R. (1983) *Intentionality. An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
- Searle, J. R. (1990). Cognitive science and the computer metaphor. In *Artificial Intelligence, Culture and Language: On Education and Work* (pp. 23-34). London: Springer London.
- Searle, J. R. (1991). Consciousness, unconsciousness and intentionality. *Philosophical Issues*, 1, 45-66.
- Searle, J. R. (2009). *The Turing test: 55 years later* (pp. 139-150). Springer Netherlands.
- Segundo-Ortin, M. (2020). Agency from a radical embodied standpoint: An ecological-enactive proposal. *Frontiers in psychology*, 11, 1319.
- Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative Artificial Intelligence: A Systematic Review and Applications. *arXiv preprint arXiv:2405.11029*.
- Senju, A., Vermetti, A., Kikuchi, Y., Akechi, H., Hasegawa, T., & Johnson, M. H. (2013). Cultural background modulates how we look at other persons' gaze. *International Journal of Behavioral Development*, 37(2), 131-136.
- Seth, A. K. (2014). The cybernetic bayesian brain. Open MIND. Frankfurt am Main: MIND Group.
- Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, 23(8-9), 1099–1103.
- Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, 16(4), 337-351.
- Shamsuddin, S., Ismail, L. I., Yussof, H., Zahari, N. I., Bahari, S., Hashim, H., & Jaffar, A. (2011). Humanoid robot NAO: Review of control and motion exploration. In *2011 IEEE international conference on Control System, Computing and Engineering* (pp. 511-516). IEEE.
- Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F. A., Mohamed, S., Piah, H. A., & Zahari, N. I. (2012). Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO. In *2012 IEEE 8th International Colloquium on Signal Processing and its Applications* (pp. 188-193). IEEE.
- Shanahan, M. (2016) The Frame Problem. *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/frame-problem/>
- Shapiro, L. & Spaulding, S. (2024). Embodied Cognition. *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/sum2024/entries/embodied-cognition/>
- Sharkey, A., & Wood, N. (2014). The Paro seal robot: demeaning or enabling. In *Proceedings of AISB* (Vol. 36, p. 2014).



- Sharkey, N. E., & Ziemke, T. (1998). A consideration of the biological and psychological foundations of autonomous robotics. *Connection Science*, 10(3-4), 361-391.
- Sharkey, N. E., & Ziemke, T. (2000). Life, Mind, and Robots. The Ins and Outs of Embodied Cognition. In S. Wermter and R. Sun (Eds.), *Hybrid Neural Systems*. Berlin: Springer, 313-332.
- Sharkey, N. E., & Ziemke, T. (2001). Mechanistic versus phenomenal embodiment: Can robot embodiment lead to strong AI?. *Cognitive Systems Research*, 2(4), 251-262.
- Shaw, G. B. (1913) *Pygmalion*. (tr. it. *Pigmalione*, Mondadori, Milano, 1980, e-book).
- Shea, N. (2018). *Representation in cognitive science*. Cambridge, MA: Oxford University Press.
- Shelley, M. (1818) *Frankenstein; or, The Modern Prometheus* (tr. it. *Frankenstein, o il modern Prometeo*, Milano, Feltrinelli, 2013).
- Sheridan, T. B. (2016). Human–Robot Interaction: Status and Challenges. *Human Factors*, 58(4), 525-532. <https://doi.org/10.1177/0018720816644364>
- Shin, H. I., & Kim, J. (2020). My computer is more thoughtful than you: Loneliness, anthropomorphism and dehumanization. *Current Psychology*, 39, 445-453.
- Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair! An interaction with a cheating robot. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 219–226)
- Shortliffe, E. H. (1977, October). Mycin: A knowledge-based computer program applied to infectious diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (p. 66). American Medical Informatics Association.
- Sierra Rativa, A., Postma, M., & van Zaanen, M. (2020). The uncanny valley of the virtual (animal) robot. In *Robotics in Education: Current Research and Innovations 10* (pp. 419-427). Springer International Publishing.
- Sierra Rativa, A., Postma, M., & van Zaanen, M. (2022). The uncanny valley of a virtual animal. *Computer Animation and Virtual Worlds*, 33(2), e2043.
- Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, 105(2), 809-813.
- Simon, H. (1969) *The Sciences of the Artificial*. Cambridge, MA: Cambridge University Press.
- Simondon, G. (1958). *Du mode d'existence des objets techniques*. Presses Universitaires de France.
- Sini, C. (2009) *L'uomo, la macchina, l'automa. Lavoro e conoscenza tra futuro prossimo e passato remoto*. Torino: Bollati Boringhieri.
- Sinigaglia, C. (2009). Mirror in action. *Journal of Consciousness Studies*, 16(6-7), 309-334.
- Smithers, T. (1997). Autonomy in robots and other agents. *Brain and cognition*, 34(1), 88-106.
- Söderlund, M., & Oikarinen, E. L. (2021). Service encounters with virtual agents: an examination of perceived humanness as a source of customer satisfaction. *European Journal of Marketing*, 55(13), 94-121.

- Somenzi, V. (1991) *La materia pensante*. Milano: Metope.
- Sone, Y. (2016). *Japanese robot culture*. Palgrave Macmillan US.
- Sorge, F. (2019) *Preistoria robotica. Dai congegni ellenistici agli autoreplicanti. Lo straordinario viaggio dell'uomo verso le machine per fare e per pensare*. Firenze: Olimpia.
- Spatola, N. (2019). Switch off a Robot, Switch off a Mind?. In *Proceedings of the 7th international conference on human-agent interaction* (pp. 194-199).
- Spatola, N. & Wudarczyk, O. A. (2021). Implicit attitudes towards robots predict explicit attitudes, semantic distance between robots and humans, anthropomorphism, and prosocial behavior: From attitudes to human–robot interaction. *International Journal of Social Robotics*, 13(5), 1149-1159.
- Spatola, N. & Wykowska. (2021). The Personality of Anthropomorphism: How the Need for Cognition and the Need for Closure Define Attitudes and Anthropomorphic Attributions toward Robots. *Comput. Hum. Behav.* 122, 106841. doi:10.1016/j.chb.2021.106841
- Spatola, N., Kühnlenz, B., & Cheng, G. (2021d). Perception and evaluation in human–robot interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A multicomponent approach of anthropomorphism. *International Journal of Social Robotics*, 13(7), 1517-1539.
- Spatola, N., Marchesi, S., & Wykowska, A. (2021a). Cross-cultural framework of anthropomorphism toward robots. Differences across cultures and ontological limits in current frameworks. *Preprint at <https://psyarxiv.com/4x92c>*.
- Spatola, N., Marchesi, S., & Wykowska, A. (2021b). The intentional stance test-2: How to measure the tendency to adopt intentional stance towards robots. *Frontiers in Robotics and AI*, 8, 666586. doi: 10.3389/frobt.2021.666586
- Spatola, N., Marchesi, S., & Wykowska, A. (2022). Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism. *Frontiers in Robotics and AI*, 9, 863319.
- Spatola, N., Normand, J. M., & Chaminade, T. (2021c). When the mind is embodied: Measuring bodily reactions to others' robot interactions. *Cognition*, 211, 104636. <https://doi.org/10.1016/j.cognition.2021.104636>
- Spatola, N., Wudarczyk, O. A., Normand, J.-M., & Chaminade, T. (2018). Not as bad as it seems: The absence of an uncanny valley for social robots with real-time social capabilities. *Computers in Human Behavior*, 88, 200-205.
- Spaulding, S. (2018). *How we understand others: Philosophy and social cognition*. Routledge.
- Spaulding, S. (2020). What is mindreading?. *Wiley Interdiscip Rev Cogn Sci*. 2020 May;11(3):e1523. doi: 10.1002/wcs.1523. Epub 2019 Dec 17. PMID: 31849192.
- Spreng, R. N., & Andrews-Hanna, J. R. (2015). The default network and social cognition. In Toga, A. (Ed.), *Brain mapping: An encyclopedic reference*, 3(December 2015), 165-169.



- Spreng, R. N., Stevens, W. D., Viviano, J. D., & Schacter, D. L. (2016). Attenuated anticorrelation between the default and dorsal attention networks with aging: evidence from task and rest. *Neurobiology of aging*, 45, 149-160.
- Springer, K., Meier, J. A., & Berry, D. S. (1996). Nonverbal bases of social perception: Developmental change in sensitivity to patterns of motion that reveal interpersonal events. *Journal of Nonverbal Behavior*, 20, 199-211.
- Stano, P., Kuruma, Y., & Damiano, L. (2018). Synthetic biology and (embodied) artificial intelligence: opportunities and challenges. *Adaptive Behavior*, 26(1), 41-44.
- Stafford, R. Q., MacDonald, B. A., Jayawardena, C., Wegner, D. M., & Broadbent, E. (2014). Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. *International journal of social robotics*, 6, 17-32.
- Stawarska, B. (2010). Mutual gaze and intersubjectivity. *Handbook of phenomenology and cognitive science*, 269-282.
- Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences*, 106(43), 18362-18366.
- Steels, L. (1990) Toward a theory of emergent functionality. In Meyer, J.-A., & Wilson, S. (Eds.) *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*
- Steels, L. (1995) Building agents out of autonomous systems. In L. Steels and Brooks (Eds.) *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale, N.J., Erlbaum, pp. 83-121.
- Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, 160, 43-50.
- Steiner, P. (2019). Untangling the Knot of Intentionality: Between Directedness, Reference, and Content. *Studia Semiotyczne*, 33(1), 83-104.
- Steiner, P. (2023). The structure of intentionality: insights and challenges for enactivism. *Inquiry*, 1-32.
- Steiner, P. (2025). Not thinking about the same thing. Enactivism, pragmatism and intentionality. *Phenomenology and Cognitive Science* 24, 9–32. <https://doi.org/10.1007/s11097-023-09949-4>
- Steinfeld, O., Jenkins, C., & Scassellati, B. (2009). The oz of wizard: simulating the human for interaction research. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (HRI '09)*. Association for Computing Machinery, New York, NY, USA, 101–108. <https://doi.org/10.1145/1514095.1514115>
- Stenzel, A., Chinellato, E., Bou, M. A. T., Del Pobil, Á. P., Lappe, M., & Liepelt, R. (2012). When humanoid robots become human-like interaction partners: corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance*, 38(5), 1073.

- Stevenson, R. L. (1886) *The Strange Case of Dr Jekyll and Mr Hyde* (tr. it. *Lo strano caso del dottor Jekyll e mister Hyde*, Milano, Mondadori).
- Stewart, J., Gapenne, O. & Di Paolo, E. (Eds.) (2010). *Enaction: Toward a new paradigm for Cognitive Science*. MIT Press.
- Stix, G. (2008). Into the uncanny valley. *Scientific American*, 299(6), 24-28.
- Stone, W. L. (2018). The history of robotics. In Kurfess, T. (Ed.) *Robotics and automation handbook*. CRC Press, 8-19.
- Strait, M. K., Floerke, V. A., Ju, W., Maddox, K., Remedios, J. D., Jung, M. F., & Urry, H. L. (2017). Understanding the uncanny: Both atypical features and category ambiguity provoke aversion toward humanlike robots. *Frontiers in psychology*, 8, 1366.
- Strait, M., Urry, H. L., & Muentener, P. (2019). Children's responding to humanlike agents reflects an uncanny valley. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 506-515). IEEE.
- Strait, M., Vujovic, L., Floerke, V., Scheutz, M., & Urry, H. (2015). Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3593-3602).
- Straub, I. (2016). "It looks like a human!" The interrelation of social presence, interaction, and agency ascription: A case study about the effects of an android robot on social agency ascription. *AI & Society*, 31(4), 553–571.
- Sturgeon, S., Palmer, A., Blankenburg, J., & Feil-Seifer, D. (2019, October). Perception of social intelligence in robots performing false-belief tasks. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1-7). IEEE.
- Sun, R. (2004). Desiderata for cognitive architectures. *Philosophical psychology*, 17(3), 341-373.
- Sutherland, C. A., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science*, 8(5), 538-551.
- Taieb, H. (2018). Introduction: Is Intentionality a Relation?. In Taieb, H. (Ed), *Relational Intentionality: Brentano and the Aristotelian Tradition*, Cham: Springer, 1-13.
- Takahashi, H., Ban, M., & Asada, M. (2016). Semantic differential scale method can reveal multi-dimensional aspects of mind perception. *Frontiers in Psychology*, 7, 1717
- Takahashi, H., Saito, C., Okada, H., & Omori, T. (2013). An investigation of social factors related to online mentalizing in a human-robot competitive game. *Japanese Psychological Research*, 55(2), 144–153.
- Tamburrini, G., & Datteri, E. (2005). Machine experiments and theoretical modelling: From cybernetic methodology to neuro-robotics. *Minds and Machines*, 15, 335-358.
- Tan, H., Wang, D., & Sabanovic, S. (2018). Projecting life onto robots: The effects of cultural factors and design type on multi-level evaluations of robot anthropomorphism. In *Proceedings of the*

2018 27th IEEE International Symposium on Robot and Human Interactive Communication (pp. 129–136).

- Tanney, J. (2022). Gilbert Ryle. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition). Retrieved from <https://plato.stanford.edu/archives/sum2022/entries/ryle/>
- Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., ... & David, D. O. (2012). Children with autism social engagement in interaction with Nao, an imitative robot: A series of single case experiments. *Interaction studies*, 13(3), 315-347.
- Telotte, J. P. (2016). *Robot ecology and the science fiction film*. New York: Routledge.
- Terada, K., & Yamada, S. (2017). Mind-reading and behavior-reading against agents with and without anthropomorphic features in a competitive situation. *Frontiers in Psychology*, 8, 1071.
- Terada, K., Shamoto, T., Ito, A., & Mei, H. (2007). Reactive movements of non-humanoid robots cause intention attribution in humans. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 3715–3720)
- Thalmann, N. M., Mishra, N., & Tulsulkar, G. (2021). Nadine the Social Robot: Three Case Studies in Everyday Life. In *Social Robotics: 13th International Conference, ICSR 2021, Singapore, Singapore, November 10–13, 2021, Proceedings 13* (pp. 107-116). Springer International Publishing.
- Thelen, M. H., Miller, D. J., Fehrenbach, P. A., Frautschi, N. M., & Fishbein, M. D. (1980). Imitation during play as a means of social influence. *Child Development*, 918-920.
- Thellman, S. (2021). *Social robots as intentional agents* (Doctoral dissertation, Linköping University Electronic Press).
- Thellman, S., & Ziemke, T. (2019). The intentional stance toward robots: Conceptual and methodological considerations. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1097–1103). Cognitive Science Society.
- Thellman, S., & Ziemke, T. (2020). Do you see what I see? Tracking the perceptual beliefs of robots. *Iscience*, 23(10).
- Thellman, S., & Ziemke, T. (2021). The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(3), 1-15.
- Thellman, S., De Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4), 1-51.
- Thellman, S., Giagtzidou, A., Silvervarg, A., & Ziemke, T. (2020). An implicit, non-verbal measure of belief attribution to robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 473-475).
- Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. Humanoid robot behavior: exploring the intentional stance toward robots. *Frontiers in Psychology*, 8 (NOV), 1-14.

- Thill, S., & Ziemke, T. (2015). Interaction as a bridge between cognition and robotics. In *Cognition: a bridge between robotics and interaction workshop in conjunction with HRI2015, 10th ACM/IEEE International Conference on Human-Robot Interaction, Portland, USA, March 2-5, 2015*.
- Thom, R. (1975). *Structural stability and morphogenesis*. Benjamin Inc., Reading.
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of the mind*. Harvard University Press.
- Thompson, E., & Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, 28, 23-30.
- Thompson, J. C., Trafton, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception*, 40(6), 695-704.
- Tinwell, A. (2014). *The uncanny valley in games and animation*. CRC press.
- Tinwell, A., & Sloan, R. J. (2014). Children's perception of uncanny human-like virtual characters. *Computers in Human Behavior*, 36, 286-296.
- Tinwell, A., Grimshaw, M., & Nabi, D. A. (2015). The effect of onset asynchrony in audio-visual speech and the Uncanny Valley in virtual characters. *International Journal of Mechanisms and Robotic Systems*, 2(2), 97-110.
- Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human behavior*, 27(2), 741-749.
- Tinwell, A., Nabi, D. A., & Charlton, J. P. (2013). Perception of psychopathy and the Uncanny Valley in virtual characters. *Computers in Human Behavior*, 29(4), 1617-1625.
- Todo, T., & Miisho, T. (2014, October). Interactions on eyeballs of humanoid-robots. In *Proceedings of the second international conference on Human-agent interaction* (pp. 281-283).
- Togai, M. (1984). Japan's next generation of robots. *Computer*, 17(03), 19-25.
- Tomasello, M. (1998). Social cognition and the evolution of culture. In *Piaget, evolution, and development* (pp. 229-254). Psychology Press.
- Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, 115(34), 8491-8498.
- Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental science*, 10(1), 121-125.
- Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child development*, 1454-1463.
- Tondu, B. (2015). Fear of the death and uncanny valley: A Freudian perspective. *Interaction Studies*, 16(2), 200-205.

- Torrance, S., & Froese, T. (2011). An inter-enactive approach to agency: participatory sense-making, dynamics, and sociality. *Humana. Mente*, 15(2011), 21-53.
- Trevarthen, C. B. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullock (Ed.), *Before Speech*. Cambridge, MA: Cambridge University Press, pp. 321-47.
- Trevarthen, C. B. (1998). The concept and foundations of infant intersubjectivity. In S. Bråten (Ed.), *Intersubjective communication and emotion in early ontogeny* (pp. 15-46). Cambridge University Press.
- Tronick, E. Z., & Cohn, J. F. (1989). Infant-mother face-to-face interaction: Age and gender differences in coordination and the occurrence of miscoordination. *Child development*, 85-92.
- Trovato, G., & Eyssel, F. (2017). Mind attribution to androids: A comparative study with Italian and Japanese adolescents. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 561-566)
- Tschöpe, N., Reiser, J. E., & Oehl, M. (2017, March). Exploring the uncanny valley effect in social robotics. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 307-308).
- Tu, Y. C., Chien, S. E., & Yeh, S. L. (2020). Age-related differences in the uncanny valley effect. *Gerontology*, 66(4), 382-392.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Journal of Mathematics*, 58(345-363), 5.
- Turing, A. M. (1950) Computing Machinery and Intelligence. *Mind*, 59: 433-460.
- Turkle, S. (1984). *The second self: Computers and the human spirit*. MIT Press.
- Turkle, S. (1995) *Life on the Screen. Identity in the Age of Internet*. New York: Simon & Schuster.
- Turkle, S. (2011) *Alone Together: why we expect more from technology and less from each other?*. New York: Basic Books.
- Turkle, S. (Ed.) (2007) *Evocative Objects: Things We Think With*. MIT Press.
- Umbrello, S. (2021). AI Winter. In Michael Klein & Philip Frana (eds.), *Encyclopedia of Artificial Intelligence: The Past, Present, and Future of AI*. Abc-Clio. pp. 7-8.
- Uono, S., & Hietanen, J. K. (2015). Eye contact perception in the west and east: A cross-cultural study. *Plos one*, 10(2), e0118094.
- Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia*, 114, 181-185.
- Urgen, B. A., Li, A. X., Berka, C., Kutas, M., Ishiguro, H., & Saygin, A. P. (2015). Predictive coding and the Uncanny Valley hypothesis: Evidence from electrical brain activity. *Cognition: a bridge between robotics and interaction*, 15-21.

- Urgen, B. A., Plank, M., Ishiguro, H., Poizner, H., & Saygin, A. P. (2013). EEG theta and Mu oscillations during perception of human and robot actions. *Frontiers in Neurorobotics*, 7, 19. <https://doi.org/10.3389/fnbot.2013.00019>
- Urquiza-Haas, E. G., & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal behaviour*, 109, 167-176.
- Vaitonytė, J., Alimardani, M., & Louwerse, M. M. (2023). Scoping review of the neural evidence on the uncanny valley. *Computers in Human Behavior Reports*, 9, 100263.
- Valagkouti, I. A., Troussas, C., Krouska, A., Feidakis, M., & Sgouropoulou, C. (2022). Emotion recognition in human–robot interaction using the NAO robot. *Computers*, 11(5), 72.
- Van Oost, E., & Reed, D. (2011). Towards a sociological understanding of robots as companions. In *Human-Robot Personal Relationships: Third International Conference, HRPR 2010, Leiden, The Netherlands, June 23-24, 2010, Revised Selected Papers 3* (pp. 11-18). Springer Berlin Heidelberg.
- Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, 48(3), 564–584.
- Varela, F. J. (1979). *Principles of Biological Autonomy*. New York: Elsevier/North-Holland.
- Varela, F. J. (1988). Structural coupling and the origin of meaning in a simple cellular automation. In *The semiotics of cellular communication in the immune system* (pp. 151-161). Springer Berlin Heidelberg.
- Varela, F. J. (1992a). Autopoiesis and a biology of intentionality. In *Proceedings of the Workshop "Autopoiesis and Perception"*, edited by B. McMullin, 4–14. Dublin: Dublin City University.
- Varela, F. J. (1992b). Making it concrete: Before, during, and after breakdowns. In Ogilvy, J. (Ed.) *Revisioning Philosophy*, Albany: State University of New York Press, pp. 97-109.
- Varela, F. J., Thompson, E., & Rosch, E. (1991) *The Embodied Mind. Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press (tr. it. *La mente nel corpo. Scienze cognitive ed esperienza umana*. Roma, Astrolabio, 2024).
- Varga, S. (2015). Intentionality, normativity and naturalism. *Philosophia*, 43(3), 611-624.
- Veit, W. (2022). Revisiting the Intentionality All-Stars. *Review of Analytic Philosophy*, 2(1), 1-24.
- Vélez, P., Gallegos, K., Silva, J., Tumalli, L., and Vaca, C. 2014. ROBSNA: Social Robot for Interaction and Learning Therapies. In A. Natraj, S. Cameron, C. Melhuish, & M. Witkowski, *Towards Autonomous Robotic Systems* (Vol. 8069, pp. 458-469): Springer Berlin Heidelberg.
- Veling, L. (2017). Becoming Real: An Anthropological Approach to Evaluating Robots in the Real World. In: Kheddar, A., *et al.* Social Robotics. ICSR 2017. Lecture Notes in Computer Science, vol 10652. Springer, Cham. [https://doi.org/10.1007/978-3-319-70022-9\\_63](https://doi.org/10.1007/978-3-319-70022-9_63)
- Veling, L., & McGinn, C. (2021). Qualitative research in HRI: A review and taxonomy. *International Journal of Social Robotics*, 13, 1689-1709.
- Vernon, D. (2014). *Artificial cognitive systems: A primer*. Cambridge, MA: MIT Press.

- Vernon, D. (2022). Cognitive Architectures. In Cangelosi, A. & Asada, M. (Eds.) *Cognitive Robotics*, MIT Press, pp. 191-212.
- Vernon, D., Metta, G., & Sandini, G. (2007). The iCub Cognitive Architecture: Interactive Development in a Humanoid Robot. In *2007 IEEE 6th International Conference on Development and Learning*, 122–127. New York: IEEE.
- Vernon, D., von Hofsten, C., & Fadiga, L. (2016). Desiderata for Developmental Cognitive Architectures. *Biologically Inspired Cognitive Architectures*, 18:116–127.
- Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., & Sandini, G. (2016b). Human activity detection on the robot iCub. [https://www.researchgate.net/profile/Nicoletta-Noceti/publication/308721175\\_Human\\_activity\\_detection\\_on\\_the\\_robot\\_iCub/links/57ecd42e08ae92eb4d266cca/Human-activity-detection-on-the-robot-iCub.pdf](https://www.researchgate.net/profile/Nicoletta-Noceti/publication/308721175_Human_activity_detection_on_the_robot_iCub/links/57ecd42e08ae92eb4d266cca/Human-activity-detection-on-the-robot-iCub.pdf)
- Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., & Sandini, G. (2017). Detecting biological motion for human–robot interaction: A link between perception and action. *Frontiers in Robotics and AI*, 4, 14.
- Vignolo, A., Noceti, N., Sciutti, A., Rea, F., Odone, F., & Sandini, G. (2016a). The complexity of biological motion. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 66-71). IEEE.
- Vignolo, A., Rea, F., Noceti, N., Sciutti, A., Odone, F., & Sandini, G. (2016c). Biological movement detector enhances the attentive skills of humanoid robot iCub. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)* (pp. 338-344). IEEE.
- Violi, P. (1997). *Significato ed esperienza*. Milano: Bompiani.
- Violi, P. (2012). How our bodies become us: Embodiment, semiosis and intersubjectivity. *Cognitive Semiotics*, 4(1), 57-75.
- Viviani, P., & Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1), 32.
- Völlm, B. A., Taylor, A. N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., ... & Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage*, 29(1), 90-98.
- Von Der Pütten, A. M., & Krämer, N. C. (2012). A survey on robot appearances. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*(pp. 267-268).
- von Zitzewitz, J., Boesch, P.M., Wolf, P. et al. Quantifying the Human Likeness of a Humanoid Robot. *Int J of Soc Robotics* 5, 263–276 (2013). <https://doi.org/10.1007/s12369-012-0177-4>
- Voss, L. (2021). *More Than Machines? The Attribution of (In)Animacy to Robot Technology*. Bielenfeld: transcription Verlag.
- Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Kaouri, C., Woods, S., ... & Werry, I. (2005, August). The influence of subjects' personality traits on personal spatial zones in a



- human-robot interaction experiment. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* (pp. 347-352). IEEE.
- Walters, M. L., Koay, K. L., Syrdal, D. S., Campbell, A., & Dautenhahn, K. (2013, August). Companion robots for elderly people: Using theatre to investigate potential users' views. In *2013 IEEE RO-MAN* (pp. 691-696).
- Wang, S., & Rochat, P. (2017). Human perception of animacy in light of the uncanny valley phenomenon. *Perception*, *46*(12), 1386-1411.
- Wang, S., Cheong, Y. F., Dilks, D. D., & Rochat, P. (2020). The uncanny valley phenomenon and the temporal dynamics of face animacy perception. *Perception*, *49*(10), 1069-1089.
- Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, *19*(4), 393-407.
- Wang, X., & Krumhuber, E. G. (2018). Mind perception of robots varies with their economic versus social function. *Frontiers in psychology*, *9*, 1230.
- Wang, Z., Deng, C., Liu, C., Jiang, Z., Geng, H., Xu, H., ... & Zhu, Y. (2025) Generative Models for Robot Learning. In *ICLR 2025 Workshop Proposals*.
- Waytz, A., & Epley, N. (2012). Social connection enables dehumanization. *Journal of experimental social psychology*, *48*(1), 70-76.
- Waytz, A., Cacioppo, J., & Epley, N. (2010b). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219-232.
- Waytz, A., Epley, N., & Cacioppo, J. T. (2010a). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*(1), 58-62.
- Weinberg, M. K., Beeghly, M., Olson, K. L., & Tronick, E. (2008). A Still-face Paradigm for Young Children: 2½ Year-olds' Reactions to Maternal Unavailability during the Still-face. *The journal of developmental processes*, *3*(1), 4-22.
- Weiss, A. & Bartneck, C. (2015). Meta analysis of the usage of the Godspeed Questionnaire Series. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Kobe, Japan, 2015, pp. 381-388, doi: 10.1109/ROMAN.2015.7333568.
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman & Co.
- Wellman, H. M. (2010). Developing a theory of mind. In Goswami, U. (Ed), *The Wiley-Blackwell handbook of childhood cognitive development*, *2*, 258-284.
- Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false-belief. *Child Development*, *72*(3), 655-684.



- West, D. M., & Travis, L. E. (1991). The computational metaphor and artificial intelligence: A reflective examination of a theoretical falsework. *AI magazine*, 12(1), 64-64.
- Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *International journal of philosophical studies*, 16(3), 323-349.
- Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron*, 40(3), 655-664.
- Wiener, N. (1948) *Cybernetics, or Control and Communication in the Animal and in the Machine*. Cambridge, MA: The MIT Press (tr. it. *La cibernetica*, Milano, il Saggiatore, 1968).
- Wiener, N. (1950) *The Human Use of Human Beings*. Boston: Houghton Mifflin Co. (tr. it. *Introduzione alla cibernetica. L'uso umano degli esseri umani*, Torino, Bollati Boringhieri, 1866).
- Wiener, N. (1966). *God & Golem, Inc.: A comment on certain points where cybernetics impinges on religion* (Vol. 42). MIT press. (tr. it. *Dio & Golem S.p.a. Cibernetica e religione*, Torino, Bollati Boringhieri, 1991).
- Wiese, E., & Weis, P. P. (2020). It matters to me if you are human-examining categorical perception in human and nonhuman agents. *International Journal of Human-Computer Studies*, 133, 1-12.
- Wiese, E., Mandell, A., Shaw, T., & Smith, M. (2019). Implicit mind perception alters vigilance performance because of cognitive conflict processing. *Journal of Experimental Psychology: Applied*, 25(1), 25-40
- Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology*, 8, 281017.
- Wiese, E., Weis, P. P., Bigman, Y., Kapsaskis, K., & Gray, K. (2021). It's a match: Task assignment in human-robot collaboration depends on mind perception. *International Journal of Social Robotics*, 14(1), 141-148.
- Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PLoS ONE*, 7(9), e45391.
- Winerman L. (2006, February). *The culture-cognition connection: Recent research suggests that Westerners and East Asians see the world differently-literally*. Retrieved from <http://www.apa.org/monitor/feb06/connection.aspx>
- Winfield, A. F. (2018). Experiments in artificial theory of mind: From safety to storytelling. *Frontiers in Robotics and AI*, 5, 75.
- Winograd, T. (1972) *Understanding Natural Language*. Edinburgh: Edinburgh University Press.
- Winograd, T. (1973). A procedural model of language understanding. In Schank, R. & Colby, K. (Eds.) *Computer Models of Thought and Language*, San Francisco, W. H. Freeman, 152-186.
- Winograd, T. (1976) Artificial Intelligence and Language Comprehension, In Winograd, T., Goldstein, I., Papert, S., Minsky, M., & Wilks, Y. (Eds.) *Artificial Intelligence and Language Comprehension*, Washington, National Institute of Education, 1-26.

- Winograd, T. & Flores, F. (1986) *Understanding Computers and Cognition. A New Foundation for Design*. Norwood, NJ: Ablex.
- Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell Publishing (tr. it. *Ricerche filosofiche*, Einaudi, 2009).
- Wood, G. (2002) *Living dolls. A magical history of the quest for mechanical life*. London: Faber and Faber.
- Woods, S., Walters, M., Koay, K. L., & Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach. In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control* (pp. 750-755). IEEE. <https://doi.org/10.1109/AMC.2006.1631754>
- Woods, V., & Hare, B. (2019). Uncanny valley of the apes. In *Why We Love and Exploit Animals* (pp. 104-120). Routledge.
- Wright, J., Sanders, T., & Hancock, P. A. (2013). Identifying the role of attributions in human perceptions of robots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 1288-1292). Sage CA: Los Angeles, CA: SAGE Publications.
- Wykowska, A. (2021). Robots as mirrors of the human mind. *Current Directions on Psychological Science*, 30, 34-40. doi: <https://doi.org/10.1177/0963721420978609>
- Wykowska, A. (Ed.) (2024). *Intentional Stance Towards Humanoid Robots. Lessons Learned from Studies in Human-Robot Interaction*. Springer: Cham.
- Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents: What can we learn from priming human social cognition? *Frontiers in Psychology*, 7, 1267.
- Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Althaus, P., & Cheng, G. (2015). "Humans are well tuned to detecting agents among non-agents: Examining the perceptual
- Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *Cognition*, 130(1), 128–134.
- Xu, K., Chen, X., Liu, F., & Huang, L. (2024). What did you hear and what did you see? Understanding the transparency of facial recognition and speech recognition systems during human–robot interaction. *New Media & Society*, 14614448241256899.
- Yam, K. C., Bigman, Y., & Gray, K. (2021). Reducing the uncanny valley by dehumanizing humanoid robots. *Computers in Human Behavior*, 125, 106945.
- Yam, K. C., Tan, T., Jackson, J. C., Shariff, A., & Gray, K. (2023). Cultural Differences in People's Reactions and Applications of Robots, Algorithms, and Artificial Intelligence. *Management and Organization Review*, 19(5), 859-875.
- Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the “uncanny valley” phenomenon. *Japanese psychological research*, 55(1), 20-32.

- Yamaji, Y., Miyake, T., Yoshiike, Y., De Silva, P. R. S., & Okada, M. (2010). STB: Intentional stance grounded child-dependent robot. In *Proceedings of the International Conference on Social Robotics* (pp. 71–80).
- Yamamoto, K., Tanaka, S., Kobayashi, H., Kozima, H., & Hashiya, K. (2009). A non-humanoid robot in the “uncanny valley”: experimental analysis of the reaction to behavioral contingency in 2–3 year old children. *PloS one*, 4(9), e6974.
- Yamazaki, A., Yamazaki, K., Burdelski, M., Kuno, Y., & Fukushima, M. (2010). Coordination of verbal and non-verbal actions in human–robot interaction at museums and exhibitions. *Journal of Pragmatics*, 42(9), 2398-2414.
- Yanco, H. A., & Drury, J. L. (2002). A taxonomy for human-robot interaction. In *Proceedings of the AAAI fall symposium on human-robot interaction* (pp. 111-119).
- Yanco, H. A., & Drury, J. L. (2004). Classifying human-robot interaction: an updated taxonomy. In *2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583)* (Vol. 3, pp. 2841-2846). IEEE.
- Ye, S., Neville, G., Schrum, M., Gombolay, M., Chernova, S., & Howard, A. (2019). Human trust after robot mistakes: Study of the effects of different forms of robot communication. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1-7). IEEE.
- Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nature reviews neuroscience*, 22(3), 181-192.
- Yin, J., Wang, S., Guo, W., & Shao, M. (2021). More than appearance: the uncanny valley effect changes with a robot’s mental capacity. *Current Psychology*, 1-12.
- Yilmaz, E., Fal, M., & Acartürk, C. (2020). The role of gaze as a deictic cue in human robot interaction. In *Augmented Cognition. Human Cognition and Behavior: 14th International Conference, AC 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22* (pp. 466-478). Springer International Publishing.
- Yuan, K., Sajid, N., Friston, K., & Li, Z. (2023). Hierarchical generative modelling for autonomous robots. *Nature Machine Intelligence*, 5(12), 1402-1414.
- Zahavi, D. (2005). Intentionality and experience. *Synthesis Philosophica*, 20(2), 299-318.
- Zahavi, D. (2010). Naturalized phenomenology. *Handbook of phenomenology and cognitive science*, 2-19.
- Zahavi, D. (2014). *Self and other: Exploring subjectivity, empathy, and shame*. Oxford University Press, USA.
- Zahavi, D. (2021). We in me or me in we? Collective intentionality and selfhood. *Journal of Social Ontology*, 7(1), 1-20.
- Zawidzki, T. W. (2015). Dennett’s strategy for naturalizing intentionality: An innovative play at second base. *Philosophia*, 43, 593-609.

- Zellini, P. (1980). *Breve storia dell'infinito*. Milano: Adelphi.
- Zhang, J., Li, S., Zhang, J. Y., Du, F., Qi, Y., & Liu, X. (2020). A literature review of the research on the uncanny valley. In *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments: 12th International Conference, CCD 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I* 22 (pp. 255-268). Springer International Publishing.
- Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition*, 224, 105076.
- Zhao, X., Cusimano, C., & Malle, B. F. (2016). Do people spontaneously take a robot's visual perspective? In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 335–342).
- Zheng, S. (2017, April 24). China's 'robot goddess' Jiajia fluffs live interview in English. South China Morning Post. Retrieved from <https://www.scmp.com/news/china/society/article/2090112/chinas-humanoid-robot-jia-jia-tongue-tied-interview-us-journalist>.
- Ziemke, D., & Bischoff, J. (2023). Accessibilities by shared autonomous vehicles under different regulatory scenarios. *Procedia Computer Science*, 220, 747-754.
- Ziemke, T. (1999). Rethinking Grounding. In: Riegler, A., Peschl, M., von Stein, A. (eds) *Understanding Representation in the Cognitive Sciences*. Springer, Boston, MA. [https://doi-org.ezproxy.unibo.it/10.1007/978-0-585-29605-0\\_20](https://doi-org.ezproxy.unibo.it/10.1007/978-0-585-29605-0_20)
- Ziemke, T. (2001). Are robots embodied?. In *First international workshop on epigenetic robotics Modeling Cognitive Development in Robotic Systems* (Vol. 85, pp. 701-746).
- Ziemke, T. (2003a). What's that thing called embodiment?. In *Proceedings of the 25th Annual Cognitive Science Society* (pp. 1305-1310). Psychology Press.
- Ziemke, T. (2003b). On the role of robot simulations in embodied cognitive science. *AISB journal*, 1(4), 389-99.
- Ziemke, T. (2003c). Robosemiotics and embodied enactive cognition. *Semiosis, Energy, Evolution and Development Journal*, 3, 112-124.
- Ziemke, T. (2004). Embodied AI as Science: Models of Embodied Cognition, Embodied Models of Cognition, or Both?. In: Iida, F., Pfeifer, R., Steels, L., Kuniyoshi, Y. (eds) *Embodied Artificial Intelligence. Lecture Notes in Computer Science*, vol 3139. Berlin: Springer.
- Ziemke, T. (2007). The embodied self: Theories, hunches and robot models. *Journal of Consciousness Studies*, 14(7), 167-179.
- Ziemke, T. (2016) The body of knowledge: On the role of the living body in grounding embodied cognition. *Biosystems*. 2016 Oct; 148:4-11. doi: 10.1016/j.biosystems.2016.08.005.
- Ziemke, T. (2022). Embodiment in Cognitive Science and Robotics. In Cangelosi, A. & Asada, M. (Eds.) *Cognitive Robotics*. Cambridge (MA): MIT Press, 213-231.

- Ziemke, T. (2023). Understanding social robots: attribution of intentional agency to artificial and biological bodies. *Artificial Life*, 29(3), 351-366.
- Ziemke, T., & Thill, S. (2014). Robots are not embodied! Conceptions of embodiment and their implications for social human-robot interaction. In *Sociable Robots and the Future of Social Relations* (pp. 49-53). IOS Press.
- Ziemke, T., Thill, S., & Vernon, D. (2015). Embodiment is a double-edged sword in human-robot interaction: Ascribed vs. intrinsic intentionality. In *Proceedings of the workshop on cognition: A bridge between robotics and interaction* (pp. 9-10).
- Ziemke, T., Thill, S., & Vernon, D. (2015). Embodiment is a double-edged sword in human-robot interaction: Ascribed vs. intrinsic intentionality. In *Proceedings of the workshop on cognition: A bridge between robotics and interaction* (pp. 9-10).
- Zlatev, J., Racine, T. P., Sinha, C., & Itkonen, E. (Eds.) (2008). Intersubjectivity: What makes us human. *The shared mind: Perspectives on intersubjectivity*, 114.
- Zlotowski, J., Proudfoot, D., & Bartneck, C. (2013). More Human Than Human: Does the Uncanny Curve Really Matter? In *Proceedings of the HRI2013 Workshop on Design of Humanlikeness in HRI from uncanny valley to minimal design Tokyo*, pp. 7-13.
- Zlotowski, J., Proudfoot, D., Yogeewaran, K., & Bartneck, C. (2015). Anthropomorphism: opportunities and challenges in human-robot interaction. *International journal of social robotics*, 7, 347-360.
- Zlotowski, J., Strasser, E., & Bartneck, C. (2014). Dimensions of anthropomorphism: from humanness to humanlikeness. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, HRI '14* (New York, NY: ACM), 66-73.
- Zlotowski, J., Sumioka, H., Bartneck, C., Nishio, S., & Ishiguro, H. (2017). Understanding anthropomorphism: Anthropomorphism is not a reverse process of dehumanization. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9* (pp. 618-627). Springer International Publishing.
- Zlotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics*, 10(5), 701-714.

## Filmografia

I film sono ordinati con criterio cronologico di uscita.

- 1927, *Metropolis*. Regia: Fritz Lang; Soggetto: Thea von Harbou; Sceneggiatura: Thea von Harbou; Fotografia: Karl Freund, Günther Rittau, Walter Ruttmann; Montaggio: Fritz Lang; Musica: Gottfried Huppertz (musica originale); Club Foot Orchestra, Giorgio Moroder, Edison Studio (riedizioni); Interpreti e personaggi principali: Alfred Abel (Johann (Joh) Fredersen), Rudolf Klein-Rogge (C.A. Rotwang, l'inventore), Brigitte Helm (Maria/la donna robot); Produzione: Erich Pommer; Distribuzione: Universum Film (UFA); Origine: Germania; Durata: 87 min (versione cinematografica), 153 min (director's cut).
- 1963-2009, *Astro Boy*. Origine: Giappone; Autore: Osamu Tezuka; Regia: Osamu Tezuka; Soggetto: Yoshiyuki Tomino; Sceneggiatura: Yoshiyuki Tomino; Musiche: Tatsuo Takai; Studio: Mushi Productions; Episodi: 193 (completa).
- 1968, *2001: A Space Odyssey* (tr. it. *2001: Odissea nello spazio*). Regia: Stanley Kubrick; Soggetto: Arthur C. Clarke (romanzo, *La sentinella*, 1948); Sceneggiatura: Stanley Kubrick, Arthur C. Clarke; Fotografia: Geoffrey Unsworth; Montaggio: Ray Lovejoy; Musica: AA.VV.; Interpreti e personaggi principali: Keir Dullea (David Bowman), Gary Lockwood (Frank Poole), William Sylvester (Heywood R. Floyd); Produzione: Stanley Kubrick; Distribuzione: Metro-Goldwyn-Mayer, Stanley Kubrick Productions; Origine: Stati Uniti d'America, Regno Unito; Durata: 141 min, 160 min (première cut).
- 1969-1996, *Doraemon*. Origine: Giappone. Autore: Fujiko F. Fujio; Regia: Yasuo Yamazaki (e altri); Soggetto: Fujiko F. Fujio; Sceneggiatura: Fujiko F. Fujio (e altri); Musiche: Shunsuke Kikuchi; Studio: Shin-Ei Animation; Episodi: 1.787 (inclusi episodi originali e remake).
- 1980-2018, *Star Wars* (tr. it. *Guerre stellari*). 9 produzioni. Trilogia originale: Guerre stellari (Star Wars: Episodio IV - Una nuova speranza), L'Impero colpisce ancora (Star Wars: Episodio V - L'Impero colpisce ancora), Il ritorno dello Jedi (Star Wars: Episodio VI - Il ritorno dello Jedi); Trilogia prequel: (*Star Wars: Episodio I - La minaccia fantasma*; *Star Wars: Episodio II - L'attacco dei cloni*; *Star Wars: Episodio III - La vendetta dei Sith*); Trilogia sequel: Star Wars: Il risveglio della Forza (Star Wars: Episodio VII - Il risveglio della Forza), Star Wars: Gli ultimi Jedi (Star Wars: Episodio VIII - Gli ultimi Jedi), Star Wars: L'ascesa di Skywalker (Star Wars: Episodio IX - L'ascesa di Skywalker).
- 1982, *Blade Runner*. Regia: Ridley Scott; Soggetto: Philip K. Dick (romanzo, *Il cacciatore di androidi* (Do Androids Dream of Electric Sheep?), 1968); Sceneggiatura: Hampton Fancher, David Webb Peoples; Fotografia: Jordan Cronenweth; Montaggio: Terry Rawlings (sup.), Marsha Nakashima; Musica: Vangelis; Interpreti e personaggi principali: Harrison Ford (Rick Deckard), Rutger Hauer (Roy Batty), Sean Young (Rachael), Daryl Hannah (Pris); Produzione: Michael Deeley; Distribuzione: Warner Bros.; Origine: Stati Uniti d'America, Hong Kong; Durata: 124 min (International Cut, 1982), 116 min (Director's Cut, 1992), 118 min (The Final Cut, 2007).
- 1984, *The Terminator* (tr. it. *Terminator*). Regia: James Cameron; Soggetto: James Cameron, Gale Anne Hurd; Sceneggiatura: James Cameron, Gale Anne Hurd; Fotografia: Adam Greenberg; Montaggio: Mark Goldblatt; Musica: Brad Fiedel; Interpreti e personaggi principali: Arnold Schwarzenegger (Terminator), Michael Biehn (Kyle Reese), Linda Hamilton (Sarah Connor); Produzione: Gale Anne Hurd; Distribuzione: CDE.; Origine: Stati Uniti d'America; Durata: 108 min.

- (1993-1995) *Power Rangers*, *Origine*: Giappone (La serie è stata adattata dalla serie giapponese *Kyōryū Sentai Zyuranger*, che è stata trasmessa in Giappone nel 1992); *Autore*: Haim Saban e Shuki Levy; *Regia*: Vari registi (inclusi Jonathan Tzachor e altri); *Soggetto*: Basato su *Super Sentai* di Toei Company; *Sceneggiatura*: Haim Saban, Shuki Levy e altri; *Musiche*: Saban; *Studio*: Saban Entertainment e altri; *Episodi*: 145 (nelle tre stagioni originali).
- (1995-1996) *Neon Genesis Evangelion*. *Regia*: Hideaki Anno; *Soggetto*: Hideaki Anno; *Sceneggiatura*: Hideaki Anno, Yoshiyuki Sadamoto, e altri; *Musica*: Shiro Sagisu; *Studio*: Gainax; *Episodi*: 26; *Origine*: Giappone; *Durata*: 24 min per episodio.
- 2013 *Her*, *Regia*: Spike Jonze; *Soggetto*: Spike Jonze; *Sceneggiatura*: Spike Jonze; *Fotografia*: Hoyte van Hoytema; *Montaggio*: Eric Zumbrunnen; *Musica*: Arcade Fire; *Interpreti e personaggi principali*: Joaquin Phoenix (Theodore Twombly), Scarlett Johansson (voce di Samantha), Amy Adams (Amy); *Produzione*: Annapurna Pictures; *Distribuzione*: Warner Bros.; *Origine*: USA; *Durata*: 126 min.
- 2014 *Ex Machina*, *Regia*: Alex Garland; *Soggetto*: Alex Garland; *Sceneggiatura*: Alex Garland; *Fotografia*: Rob Hardy; *Montaggio*: Mark Day; *Musica*: Ben Salisbury e Geoff Barrow; *Interpreti e personaggi principali*: Domhnall Gleeson (Caleb Smith), Alicia Vikander (Ava), Oscar Isaac (Nathan Bateman); *Produzione*: Andrew Macdonald e Allon Reich; *Distribuzione*: Universal Pictures; *Origine*: Regno Unito; *Durata*: 108 min.
- (2016-2022), *Westworld* (Serie televisiva). *Stagioni*: 4; *Episodi*: 36; *Ideatore*: Jonathan Nolan, Lisa Joy; *Soggetto*: Michael Crichton (sceneggiatore de *Il mondo dei robot*); *Interpreti e personaggi principali*: Evan Rachel Wood (Dolores Abernathy), Thandiwe Newton (Maeve Millay), Jeffrey Wright (Bernard Lowe), Anthony Hopkins (Robert Ford); *Origine*: USA; *Durata*: 50-90 min (episodio).