

Alma Mater Studiorum – Università di Bologna
in cotutela con Università del Lussemburgo

DOTTORATO DI RICERCA IN

Law, Science and Technology

Ciclo XXXIV

Settore Concorsuale: 12/H3

Settore Scientifico Disciplinare: IUS/20

**AI AND LEGAL PERSONHOOD:
A THEORETICAL SURVEY**

Presentata da: **Claudio Novelli**

Coordinatore Dottorato

Prof.ssa Monica Palmirani

Supervisore

Prof. Giovanni Sartor

Supervisore

Prof. Johan Van der Walt

Esame finale anno 2022



PhD-FDEF-2022-010
The Faculty of
Law, Economics and Finance



Alma Mater Studiorum –
Università di Bologna
in partnership with LAST-JD

DISSERTATION

Defence held on 16/06/2022 in Bologna

to obtain the degree of

DOCTOR OF THE UNIVERSITY OF BOLOGNA

IN LAW, SCIENCE AND TECHNOLOGY

AND

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN DROIT

By

CLAUDIO NOVELLI

Born on 21 December 1992 in Naples (Italy)

AI AND LEGAL PERSONHOOD: A THEORETICAL SURVEY

Dissertation defence committee:
Dr Giovanni Sartor, dissertation supervisor
Professor, Università di Bologna
Dr Johan Van der Walt, dissertation supervisor
Professor, Université du Luxembourg
Dr José Juan Moreso, chairman
Professor, Universidad Pompeu Fabra
Dr Damiano Canale
Professor, Università Bocconi
Dr Maria Cristina Redondo
Professor, Università di Genova

INDEX

INTRODUCTION	10
1. A tale of legal dogmatic.....	10
2. Working on legal personality: a regulatory and theoretical analysis 12	
2.1. Part One: Artificial Intelligence Systems in the legal domain.....	12
2.2. Part Two: a theoretical framework of legal personality	13
2.3. Terminology	15
4. Seven claims for Part One	16
5. Seven claims for Part Two.....	18
PART ONE – ARTIFICIAL INTELLIGENCE SYSTEMS IN THE LEGAL DOMAIN.....	20
I. The Legal Salience of Artificial Intelligence	21
1. The evolution of legal personhood	21
2. What is an AIs?	24
3. AI and law	28
3.1. Ceci n'est pas une chose	32
3.2. A risk-based regulatory framework for AI: the proposal of the Artificial Intelligence Act.....	35
4. Artificial minds: intelligence, autonomy and intentionality	37
4.1. Intelligence.....	38
4.2. Autonomy.....	41
4.3. Intentionality: realism and interpretivism.....	51
5. The “interaction stance”: AIs’ social skills and legal personification	
62	
6. Conclusions.	64
II. Toys or Subjects: Seeking the Legal Status of AIs	65
1. Case scenario: Herbert the trading agent.....	65

2. The use of AIs from the legal perspective: risks, damages and responsibilities	67
2.1. Accidents will happen.....	71
3. Moral responsibility, liability and accountability	74
4. How to enforce accountability for AIs: transparency and explainability (XAI)	79
4.1. Explainable AI (XAI).....	83
5. The status of AIs under criminal law.....	86
5.1. AIs with “licence to kill”	89
5.2. Unexpectedly criminal	90
5.3. Criminal personhood and direct liability.....	92
III. The Status of AIs Under Civil Law	102
1. Civil law and AI	102
1.2. Civil liability and its peculiarities	103
2. The social cost of AI	107
3. AIs as cognitive tools	110
3.1. Contract law	111
3.2. Tort law.....	116
4. AIs as (part of) juridical persons: company law proposals	122
5. AIs as individual persons in law.....	127
5.1. The limits of conventional liability models	127
5.2. The function of legal personhood on AIs.....	130
5.3. The AIs’ legal personality.....	134
5.4. The case against legal personality on AIs	143
6. Why do we need theoretical analysis of legal personality: the coordination problem	150
7. Conclusions	153
PART TWO – A THEORETICAL FRAMEWORK OF LEGAL PERSONALITY.....	155
I. The Philosophical Inquiry of Legal Personality	156
1. Introduction: philosophical tools	156

1.1. A functional and ontological account of legal personality...	158
2. Conceptual analysis and real definition.....	160
2.1 Conceptual Analysis	161
2.2. Towards a conceptual and metaphysical framework.....	165
II. A conceptual framework of legal personality	170
1. Introduction	170
2. Law's conceptual texture	171
3. Persons in the legal tradition	172
3.1. Post-Kelsenian developments and the contemporary debate	175
4. "Who is law for?"	178
5. Personality statuses and persons in law	181
7. What kind of concept is legal personality?	185
8. The nature of legal concepts I: inferentialism	189
9. The nature of legal concepts II: ontology and neo-institutionalism	194
9.1. Compositionality.....	195
9.2. Institutional reality and legal personality.....	197
9.3. A Neo-institutional account of legal personality	199
10. Developing the neo-institutional ontology: meta-institutional concepts	204
11. Conclusions	208
III. The Nature of legal personality: real definition analysis	210
1. Personality as an institutional kind: social ontology and neo-institutionalism	210
1.1. Social ontology.....	211
1.2. Social ontology and law: neo-institutionalism.....	214
1.3. What institutions are? The rules-in-equilibrium account	218
1.4. Legal institutions as social kinds	222
2. Analytic metaphysics and the structure of legal kinds.....	226
2.1. Grounding-anchoring-framing (GAF) applied to law	231

2.2. A metaphysical framework for legal kinds (and facts).....	233
2.3. The value of the grounding-anchoring diagram for legal ontology.....	236
2.4. What about normativity?	239
3. The metaphysical structure of legal personality.....	241
4. Theories of Legal Personhood: Legalism vs Realism and Pluralism vs Monism	245
5. Legal personality theories reviewed according to the anchoring-grounding diagram	249
6. Conclusion	251
IV. Combining Layers for the Scope of AIs	253
1. Conceptual and metaphysical knowledge.....	253
2. Meta-institutional <i>anchors</i> institutional?	255
3. Theoretical benefits of a multilayered ontology and intermediate legal concepts	259
4. On the existence of an intermediate (or quasi-institutional) layer: legal subjectivity	261
5. Reflective equilibrium between institutional layers: with a little help from mid-level principles	264
5.1. Back and forth between neo-institutional layers	266
6. Applying a meta-institutional and an institutional perspective to AIs	268
CONCLUSIONS AND LIMITATIONS.....	274
References	276

Abstract

I set out the pros and cons of conferring legal personhood on artificial intelligence systems (AIs), mainly under civil law. I provide functionalist arguments to justify this policy choice and identify the content that such a legal status might have. Although personhood entails holding one or more legal positions, I will focus on the distribution of liabilities arising from unpredictably illegal and harmful conduct. Conferring personhood on AIs might efficiently allocate risks and social costs, ensuring protection for victims, incentives for production, and technological innovation. I also consider other legal positions, e.g., the capacity to act, the ability to hold property, make contracts, and sue (and be sued). However, I contend that even assuming that conferring personhood on AIs finds widespread consensus, its implementation requires solving a coordination problem, determined by three asymmetries: technological, intra-legal systems, and inter-legal systems.

I address the coordination problem through conceptual analysis and metaphysical explanation. I first frame legal personhood as a node of inferential links between factual preconditions and legal effects. Yet, this inferentialist reading does not account for the ‘background reasons’, i.e., it does not explain why we group divergent situations under legal personality and how extra-legal information is integrated into it. One way to account for this background is to adopt a neo-institutional perspective and update its ontology of legal concepts with further layers: the meta-institutional and the intermediate. Under this reading, the semantic referent of legal concepts is institutional reality. So, I use notions of analytical metaphysics, such as grounding and anchoring, to explain the origins and constituent elements of legal personality as an institutional kind. Finally, I show that the integration of conceptual and metaphysical analysis can provide the toolkit for finding an equilibrium around the legal-policy choices that are involved in including (or not including) AIs among legal persons.

INTRODUCTION

1. A tale of legal dogmatic

A fictional tale often recurs in the literature on legal personality. It can be traced back to Gustav Schwarz, whose book on legal subjects – *Rechtssubjekt und Rechtszweck* (‘Legal subject and Purpose’) (Schwarz 1908) – begins with an imaginary dialogue between a group of ancient philosophers who, awakened from eternal rest, observe with amazement a scene of everyday life in a modern society (Zatti 1975). What triggers the debate among ancient philosophers is the vision of a tram on rails. Namely, they wonder what kind of vehicle it is and who steers it. They start arguing:

1st Philosopher. ‘There can be no such thing as a cart that is not pulled by a horse. Yet the strange cart I see moves forward, and no beast of burden is in sight; those who do not want to forgo an explanation can only imagine that a horse is there: it is the fictitious animal that pulls the cart’

2nd Philosopher. ‘I’m afraid you’re wrong: there’s no fiction that can move a car. Of course, a horse is necessary: but you have to look for it in reality. *Hic et nunc*, the animal is missing; it can be assumed that a horse has been in the past, or that it will be in the

future. There is no need for fiction: it is the past horses, or possibly the future ones, that pull the cart.’

3rd Philosopher: ‘You two are way off...there is no need to look for fictitious or real animals since it is clear that an organism is driving the cart, the tram company! This organism is an individual, it has everything is needed to drive a wagon: It has a head, trunk, hands and feet – that is, the management, shareholders, employees and workers. And it has a will, with which it drives the car.’

4th Philosopher: ‘All you do is talk in metaphors! With metaphorical hands and feet, you can’t even push a stroller. This cart is not driven by anyone, and even the person who sits in the front does not make any effort. The wiser thing to do is to realise that there are now horseless carts, somehow self-propelled carts’

5th Philosopher: ‘You convinced me! It is necessary to think of a different type of cart from the old one, using a different force: and not to confuse the two species’¹

At this stage our exhumed thinkers reach the conclusion that vehicles are always moved by some kind of force and it is not the animal pulling the cart that matters, but the type of force applied. In some later versions of the story, these philosophers were joined by a sixth one who urged colleagues to consider how the new driving force, whatever its nature, was still moved by men in flesh and blood. He suggested that their task was then to understand how human beings governed the new force and with what ends.

Some of the main dogmatic positions marking the debate on legal personality, from the most formalist to the most reductionist, are represented in this brief tale. When a new phenomenon catches the attention of jurists, and involves the topic of the persons' law, stances similar to those of the ancient philosophers are taken.

Contemporary jurists analysing artificial intelligence systems (AIs) resemble somewhat the ancient philosophers puzzled by the tram on

¹ This fable is for illustrative purposes only: I have translated and substantially modified the original version.

rails. Indeed, the queries that most frequently arise concern the origin of the ‘force’ that drives an AI agent, i.e. whether this origin is deterministically human, whether it is driven by an artificial spirit, or whether it is the product of their combination.

Jurists are expected to consider which is the most appropriate legal instrument to resolve the uncertainty surrounding a new phenomenon. And so metaphysical questions about the nature of artificial intelligence gradually give way to solutions which are functional to the regulatory purposes. Depending on the nature of this new entity, a number of legal questions may be addressed, some of the most relevant of which relate to accountability. In short, the time then comes to ask what it is that matters for law.

In this thesis, the theoretical hypotheses floating around the questions concerning the appropriateness and efficacy of describing artificial intelligence systems as persons in law may remind us of the positions of the six thinkers struggling with the tram puzzle.

2. Working on legal personality: a regulatory and theoretical analysis

In this thesis I carry out an analysis of legal personality and discuss whether it may be conferred on artificial intelligence systems (AIs). The thesis is divided into two parts: the first dealing with regulatory issues, the second with more distinctly philosophical ones.

2.1. Part One: Artificial Intelligence Systems in the legal domain

In Part One, I give an overview of the technical features that make AIs disruptive entities for the legal system, especially with regard to their qualification and attribution of status: are they things, subjects or do they lay the foundations for a new category of legal entity? As I shall illustrate, this is a cyclical issue in the so-called law of persons, although this typically occurs in the wake of social, ethical or economic rather than technological transformations. This is not to say that the same social, ethical or economic drives do not also play a decisive role in AIs, but that they are secondary, or triggered, by factors that are primarily computer-related. For this reason, AIs may introduce new elements to

the debate on the contours of legal personality, regardless of the regulatory choice that legislators will actually make.

A socio-technical perspective will be favoured, which means thinking and regulating AIs not as isolated systems but as parts of complex interactions involving multiple individuals and also multiple technical and organisational resources. Specifically, I intend to place the problems associated with the use of AIs in the socio-economic context by offering a reading of the consequences that these systems can create on the legal sphere of others in terms of positive and negative externalities. The instruments of law can govern economic processes through the allocation of these externalities; hopefully in the most efficient way possible.

As a matter of fact, legal personality, for which I will opt for a mainly functionalist reading, is a tool that accomplishes social and economic outcomes by allocating and combining legal positions. Among the positions relevant for these purposes, especially in reference to the use of AIs and new technologies, is that of responsibility.

In a nutshell, I describe the legally salient technical peculiarities of AIs (Chapter I), outline their social costs and the challenges to assigning liability (Chapter II), and evaluate the legal policy case for conferring legal personality on AIs. I consider both criminal and civil law personalities, but focus much more on the latter (Chapter III).

Part One concludes with the acknowledgement that the choice of attributing personalities to AIs clashes, among other things, with information asymmetries regarding both technical and legal aspects. To overcome these asymmetries, resulting in a kind of legislative coordination problem, I propose to also use conceptual tools, which I develop in the second part of the thesis.

2.2. Part Two: a theoretical framework of legal personality

In Part Two, I examine the concept of legal personality and the reality it refers to, the institutional one. I thus provide a metaphysical explanation of legal personality as an institutional kind.

After a brief methodological incipit (Chapter I), I reconstruct the *doctrinal history* of legal personality. I thus present the main theories, the assumptions from which they start, and the ways in which we can classify them. Each of these theories enhances a distinct rationale for personality, and as a result deals differently with issues related to the

possible new legal subjectivities. Some of them – mainly those that are classified as 'Realists' – tend to identify in a single feature the silver bullet triggering (or not triggering) the legal personality. Others emphasise systemic and contingent aspects – mainly the 'Legalists' – and tend to have a pluralist approach (Chapter II).

Chapter II is also devoted to *conceptual analysis* of the personality in law. First of all, the functioning of this concept is discussed, i.e. the way in which it regulates the cases subjected to it. Afterwards, I address profiles related to its semantics and referentiality: is it a mere rule of inference between factual premises and normative implications, or does it have a meaning that transcends its pragmatic dimension? And also: is there anything in reality that corresponds to the concept of personality? Finally, I try to update the neo-institutional theory of legal concepts in the light of developments on the conceptualisation of institutional practices.

Chapter III is focused on *real definition analysis* (metaphysical explanation), which aims to answer questions like: what is it for an entity to be a person of law? So, to discuss the nature of legal personality – and to account for the different philosophical assumptions that the main theories subscribe to – I use tools from metaphysics and social ontology. I suggest looking at personhood from the perspective of the legal theory most deferential to social ontology methodology: neo-institutionalism (consistently with conceptual analysis). From this angle legal personality – and other legal properties (property, marriage, wills, etc.) – can be seen as a subspecies of socio-institutional kinds or facts. Thus classical tools of social ontology – e.g., institutional facts, status functions and constitutive rules – provide a first canvas for constructing the metaphysical structure of legal personality; to these tools I propose to add those currently discussed by analytic metaphysics, e.g., supervenience, grounding and anchoring.

In the chapter on Conclusions, I try to connect the dots. Firstly, by attempting to bridge the features of conceptual and metaphysical framework, offering a (neo)institutional ontology of legal personality. Secondly, by identifying the benefits we can draw from such an ontological and conceptual approach for legal policy decisions such as whether or not to give legal personality to AI systems.

The final aim is to construct for legal concepts a multilayered ontology consisting of an institutional, an intermediate, and a meta-institutional layer. Each layer describes sets of facts, values, and

properties relevant for the personality status. The content of the institutional layer is fixed by legal rules, while the other two layers tend to be extra-legal, in a sense that I will explain. The intermediate level includes concepts that, like midlevel terms in morality, can help to generate reflexive equilibria in doctrinal and legislative debates, facilitating theoretical agreements under circumstances of indeterminacy

2.3. Terminology

A few brief clarifications on the terminology used are necessary. In the philosophical and legal literature on the law of persons, the terms 'legal personality', 'legal personhood', and 'legal capacity' are often treated as different concepts. I do not see the need for a proliferation of words expressing overlapping meanings.

Nor do I see any risk of confusing 'personality' as the psychological-aptitude profile, or 'subjectivity' as agency or consciousness, with personality in law, if these terms are preceded by the adjective 'legal' (or where there is any other reference to 'law'). To avoid excessive repetition I sometimes use legal personality and legal personhood interchangeably; the same term will be applied to individual and collective entities, and "empty" platforms (e.g. set of assets serving a specific purpose). Briefly, here is the terminology adopted in the thesis:

- *Legal personality/personhood.* The legal status with which entitlement to (at least) one legal position is officially associated: e.g., rights, duties, powers, responsibilities, powers, and immunities. What is relevant is whether the legal positions are 'active' or 'passive'; whether they are exercised 'directly' or 'indirectly'. Variations on whether some of these statuses may result in one-dimensional combinations of legal positions will be expressed through expressions such as 'mere active legal personality' and 'mere passive legal personality'; 'direct legal personality' or 'indirect legal personality'.
- *Legal subjectivity.* Mere attitude to the entitlement – or *de facto* entitlement – to legal positions. It will be portrayed as an intermediate concept, capable of mediating between the

status of personality and its extra-legal foundations, similar to what happens with mid-level principles to justify moral judgements.

I dedicate a special section to clarify what I am referring to when I talk about artificial intelligence systems (Part One, Chapter 1, section 2). This is not an easy task as this is a class of computer artefacts that may include different technologies and products, sometimes quite different from each other, e.g. an autonomous driving car and an algorithm to calculate credit worthiness. What these systems share is the inferential paradigm, mainly relying on the use of large amounts of data and statistical models to process them.

4. Seven claims for Part One

In part one, devoted to investigating the regulatory hypothesis of conferring legal personality on AIs (both under criminal and civil law), I will defend seven claims:

- (1) I argue that operational and cognitive peculiarities of AIs are relevant for the purposes of legal personality, not because they grant the ontological attributes typically grounding legal personality for individuals or groups – there is no scope for such a comparison as yet, and it is not certain that there ever will be – but rather because they pose the practical conditions for redrawing the way in which legal positions are allocated. Embedding AIs in law create new opportunities – mostly related to the advantages of allowing unsupervised action by artificial agents with very accurate decision-making capabilities – and risks – mostly related to reliability and responsibility gaps.
- (2) I propose to consider the legal profiles triggered by the deployment of AIs from a *sociotechnical* perspective. This implies that evaluations of the optimal legal strategy to regulate AIs, including their status qualification, must take into account the interplay between technical factors – e.g. unpredictability of decision-making processes due to the

interaction with data, inferential models and agents (artificial and human) – and human and organisational factors – e.g. the procedures of production and implementation of such systems, which may see the participation of multiple individuals, each with different roles and responsibilities.

- (3) The complexity of socio-technical systems precludes clear-cut solutions, which is why I sometimes follow a cost-benefit analysis approach to reasoning about the appropriateness of a legal arrangement.
- (4) I contend that one of the main reason for considering the creation of an autonomous centre of imputation of legal positions (i.e., a personality) is the management of liabilities arising from failures of AIs. However, I propose to distinguish three different failures: (f1) mistakes, (f2) misuses and (f3) accidents. While for the first two situations the conventional patterns of responsibility – which assume AIs to be cognitive products or tools – seem adequate, the third situation is the one that seems most salient to the proposal of AIs' legal personality.
- (5) Consistent with these premises, I will provide arguments supporting a functionalist justification for conferring personhood on AIs: a better distribution of liabilities resulting from unpredictably illegal and/or harmful behaviour. This requires defining what is an efficient allocation of risks and social costs associated with the use of AIs: redress for victims, incentives for production, and technological innovation. I also consider considers other legal positions triggered by personhood: competencies and powers like financial autonomy, capacity to act, the ability to hold property, make contracts, sue (and be sued).
- (6) I hypothesize what content possible status of AIs as persons in civil law might have.² The scenario of a *dependent* legal

² With respect to the personality of criminal law - to which several pages are also devoted - no structured or original claims are advanced.

personhood status that holds AIs liable for events that are not caused by misuse, negligence, or mere design errors, was considered as the most plausible and beneficial. The resulting model thus combines elements of negligence liability with elements akin to no-fault systems, with the provision of (limited) capital capacity to the AIs.

- (7) Finally, I argue that, given the political will to confer legal personality on AIs, such a decision is likely to encounter some sort of coordination problem caused by three types of asymmetries: (a1) *Technological*; (a2) *Intra-systemic*; and (a3) *Inter-systemic*.

5. Seven claims for Part Two

Part Two is devoted to metaphysical and conceptual analysis of legal personality. The theoretical purpose of this thesis is to provide a framework of the metaphysical and conceptual relationships that structure the legal kind (or fact) of personhood. I make the following claims:

- (1) From a conceptual point of view, I describe the concept of legal personality as a classic concept with clear-cut membership criteria and in hierarchical relation to the sub-types, as opposed to those who claim that it is a cluster one with fuzzy membership.
- (2) I argue that an *inferential* reading of the concept of legal personality – and related legal institutions (e.g. property or marriage) – can be defended in a weak version, but that for a more comprehensive account it is appropriate to consider a *compositional* reading of personality. I also share the idea that concepts such as that of personality have semantic referents to be located within institutional reality.
- (3) I therefore propose a conceptual ontology of legal personhood based on MacCormik's account of institutional legal concepts, but updated with other institutional layers, namely the meta-

institutional and intermediate. Both ontological layers hold implications for conceptual knowledge: at the meta-level we find concepts such as agency, self-awareness, dignity or vulnerability, at the intermediate level the concept of legal subjectivity (mere aptitude for entitlement to legal positions).

- (4) I will argue that legal personality statuses can have two conformations: (a) *thick* – in which the status is associated with a complex package of legal positions – and (b) *thin* – for which it is sufficient that there is at least one rule attributing a legal position to a given entity.
- (5) From real definition analysis perspective, I will contend that legal personality, as well as others legal categories, can be seen ontologically as a socio-institutional kind – or an institutional fact – according to MacCormick's neo-institutionalist approach, but updating it to an account that sees institutions as systems of rules in equilibrium (by Hindriks and Guala).
- (6) I use the metaphysical notions of grounding and anchoring, as developed by Epstein in his model of social ontology, to describe the constitutive (and explanatory) relations of legal kinds and facts; then I apply this model to legal personality.
- (7) I conclude by discussing the benefits of a multi-level ontology of legal of legal personality. In addition to theoretical benefits, I argue that such a conceptual and metaphysical framework can have practical ones: it can serve to find an equilibrium and build consensus around legal policy choices that are involved in including (or not including) AIs among legal persons; also providing a strategy to address the coordination problem presented above.

**PART ONE - ARTIFICIAL
INTELLIGENCE SYSTEMS IN THE
LEGAL DOMAIN**

I. The Legal Salience of Artificial Intelligence

1. The evolution of legal personhood

The notion of legal personality is a dynamic one. In the course of the different political seasons, legal systems have had to review many times the perimeter of the entities considered as proper persons in law, as well as the set of prerogatives attached to this status. The identification of the bearers of rights and duties has long been, and will be, an instrument of legal policy through which law identifies its own community. Legal personification may have important social effects, by empowering and protecting the entities personified and by working to the benefit of the general interest (e.g. by incentivising economic or social initiatives) (Dewey 1926; Hansmann, Kraakman and Squire 2005).

There is much debate about the boundaries of legal personality: about whether not only humans and their organisations, but also nonhuman animals, environmental entities, idols, unborn children, and software agents should be granted personality status (Kurki and Pietrzykowski 2017).

The contours of the status of legal personality are of major theoretical value, considering the systemic role personality plays not only within legal system, but also within other networks of rules governing human conduct (Nékam 1938).

In most legal systems, the entities just mentioned are not seen as responsible agents. Some of them are seen as incapable of purposeful

action, and others (e.g., animals), while endowed with a limited agency, are seen as incapable of reasoning or deliberating and devoid of moral standing. But it has not always been like that: there have been times and places where, for instance, nonhuman animals have been put on trial and charged with proper crimes (and even convicted) (Evans 1906).

Conversely, there have been contexts, e.g., Roman law, in which the status of a full-fledged person was only granted to certain human individuals, to the exclusion of others (e.g., slaves) (Curran 1983). Modern legal systems, on the other hand, include all human beings in the community of legal persons because only humans are deemed to be morally and mentally worthy of holding legal positions. Personality is also extended to various types of human organisations — operating in the economic, social, or religious sphere —, since such organisation advance human interests, to this end relying on the cognitive skills of their human representatives and agents.

The promoters of legal personhood for nonhuman entities, mainly in the case of animals, often claim that some of the cognitive skills typically relevant to morality and law — skills such as basic rationality, understood as the capacity to adopt goals/desires and acting (consistently) toward them — are not exclusive to humans.³ In moral philosophy, a version of this idea is also known as the *argument from marginal cases*, and it challenges the inconsistency of ascribing moral standing to marginal human beings with very limited, or no, cognitive skills —e.g., individuals who are severely mentally disabled, pre-rational (children) or post-rational (the senile) — while not doing the same for animals that meet equal or even superior rationality standards (Dombrowski 1997; Tanner 2009).

Others have argued that the capacity for purposeful action is not an essential precondition of legal personality, since personification may instead be aimed at protecting entities having properties such as sentience, dignity, or vulnerability, or at advancing interests such as preserving the environment or the welfare of the community.⁴ One objection to this view might be that legal personality is neither the only available nor the most appropriate juridical instrument with which to protect these entities or advance these interests.

³ On animal rights see, in the first place, (Singer 1975); more specifically on the autonomy of animals and legal status in (Wise, *Rattling the Cage: Towards Legal Rights for Animals* 2000).

⁴A comprehensive analysis of these profiles can be found (Gellers 2021).

These disputes rest on different understandings of the roots of legal personality, as well as of the overall legal practice around it, and reveal that the parameters used by legal systems to ascribe personality are still controversial and evolving.

Each nonhuman entity's eligibility for legal personhood deserves its own in-depth study, since different considerations may apply. For instance, the way an animal might benefit from personality is quite different from the way a river might.⁵

This thesis is mainly focused on the contemporary debate concerning the hypothesis of conferring legal personality on a peculiar type of artefact: artificial intelligence systems (AIs).

However, even among seemingly disparate entities there are similarities to be investigated (e.g. cognitive skills and operational autonomy). Having illustrated some of the reasons that make AIs potentially amenable to the granting of this status it will be stressed that the regulatory hypothesis under consideration requires a wider analysis of the concept of legal personality. I believe that the resulting framework can be effectively applied also to other entities.

Several theoretical positions regarding legal personhood and its meaning will be observed throughout this work. We shall see that there are two major views: someone think it is a matter of enhancing the essence – perhaps moral – of certain entities, while others emphasize its instrumental and systemic role. The former is a property-driven approach, the latter is a contingency-driven approach.

Even though I am more sympathetic to the latter view, as it will emerge, I think that an overly formalistic reading of legal personhood risks giving us a partial picture of legal forms, how they work and how they integrate non-legal information. I would prefer to remain as neutral as possible with regard to these two positions. Therefore, I shall submit a model of the metaphysical framework of legal personality, of the background reasons that entangle it to social and brute facts, and of the relations that occur between strictly legal and non-legal facts; trying not to indulge in any kind of reductionism. This framework is compatible with different theories of personality, i.e., with different ideas of what the metaphysical grounds and anchors of personality are.

⁵ A traditional reading here is (Stone 1972). There have also been some concrete initiatives to promote the attribution of so-called 'environmental personality', as the recent New Zealand experience proves; see here (Argyrou and Hummels 2019).

For the time being, however, I shall not engage in ontological or metaphysical questions just yet, and shall provide an overview of the entities specifically investigated in this work, AIs. I believe that the regulatory hypothesis of attributing legal personality to these entities is a case study with intriguing implications for understanding the nature and the concept of legal personality.

2. What is an AIs?

The expression ‘Artificial Intelligence systems’ (AIs) will often be used in the course of this thesis. In the next sections I shall indicate which types of technological artefacts we are covering, what are the main technical features of these systems and why they might have a disruptive impact on legal regulation to the point of evoking thoughts (philosophical or not) about the emergence of autonomous centres of interests. But first of all, let us try to give a definition of "artificial intelligence system". Having a reliable and fairly stable definition seems a key point for the legislator too.

There are many official definitions of what artificial intelligence systems are, provided both by independent organisations and by national legislators; I would take as a starting reference the one provided by the expert group appointed by the European Commission in 2019 (AI HLEG): “Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. [...] As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques

into cyber-physical systems)”.⁶ This is a very comprehensive definition, providing both a notion of AI as a field of study and of the products of its implementation, that can be taken into account for regulatory purposes.

Thus, when talking about AIs one can refer to software running either in a simple computer – or in a computer network – or in a more complex physical architecture – i.e. the hardware component - made up of e.g. sensors, computer vision devices, etc. The former might be defined as generic ‘AI algorithms’ or ‘software agents’ while the latter might be called ‘robot agents’. Either way, prior to software or hardware, the core element of both these technologies is *data* on which – at least for the new generation – the AI systems are built and trained.

For what concern robots and robotics, there are several industrial applications – especially in heavy industries – both for logistics and manufacturing. There are also interesting uses of robots for educational and recreational purposes as well as to assist vulnerable people, e.g. people suffering from dementia or autism.⁷ In addition, some robots are able to provide surgical assistance for high-precision operations.⁸

For the purpose of this thesis, it must be specified that most AI systems only perform a fraction of the activities listed in the definition by the HLEG: pattern recognition (e.g., recognising images of plants, animals, objects, human faces, or attitudes), language processing (e.g., understanding spoken languages, translating between languages, fighting spam, or answering queries), practical suggestions (e.g., recommending purchases, purveying information, performing logistic planning, or optimising industrial processes), etc. On the other hand, some systems

⁶ AI HLEG. A definition of AI: main capabilities and disciplines, European Commission, 2019: <https://ec.europa.eu/digital-inglemarket/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>.

A more concise definition is the one provided by the Organisation for Economic Co-operation and Development (OECD), in the 2019 recommendation of the Council on Artificial Intelligence: “An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy” (OECD. LEGAL/0440).

⁷ A robot assisting children suffering from autism and communication problems is (Wood, et al. 2021)

⁸ For an extensive overview of robotics applications see (Ben-Ari and Mondada 2018).

may combine many such capacities, as in the example of self-driving vehicles or military and care robots.

Since my analysis is focused on granting legal personality to AI systems, I shall only consider artificial agents that have an ability to pursue goals autonomously, meaning that while their top-level goals may be defined externally (by the agents' designers and users), the way in which such top-goals are pursued by the agents — including the choice of lower-level instrumental goals — is responsive to the physical and social environment in which the agents act. Thus, for instance, even if automated translation or a machine-learning system designed to identify objects in images qualifies as an AI system, neither would fall within the scope of our analysis. While it would make more sense to ask such a question for those robots having a higher degree of operational flexibility, social impact – e.g. capacity to affect fundamental rights – and risk tolerance. Robots of this kind are certainly employed in the automotive – one of the most discussed applications by jurists are self-driving vehicles – and in the military field, e.g. autonomous weapons and drones. Advanced digital bots used in online commerce, or physical robots providing services such as transportation or care can come within our scope as long as they process information and make choices among epistemic and practical alternatives pertaining to their tasks, make consequential changes in the digital or physical world, and engage in communication.

To understand the impact of these technologies, consider that the market for self-driving cars alone is estimated to grow by 36% per year and generate global revenues of 173 billion by 2030 (source: Market Watch). No need to repeat that these technologies constantly raises ethical and legal issues.⁹

On the other hand, if we take a look at the "non-robotic" AI algorithms, the applications are even more numerous and consistently increasing. This group includes the so-called expert systems, such as computer programmes with specialised expertise in certain areas of knowledge, e.g. medical diagnostic systems. The expert systems are problem-solving agents based on deduction rules (*if condition...then consequence*) and inference procedures. Anyway, they no longer seem to be dominant in the artificial intelligence field due to their working

⁹ Some arguments against autonomous weapons systems are in (Sharkey 2019). A more hopeful approach is taken by (Umbrello, De Bellis and Torres 2020).

stiffness and their time-consuming processes of knowledge integration. They have been replaced by more flexible technologies like artificial neural networks (ANNs) and machine learning (ML) algorithms; both technologies are based on example-driven learning methods that “allow computer programs to automatically improve through experience”.¹⁰

Just to give few examples, those ML algorithms working on already labelled training data – i.e. supervised learning – are employed for natural language processing, speech recognition, face recognition, email spam filtering, computer vision, information retrieval and computational finance. Those working on unlabelled examples – i.e. unsupervised learning – can be used for finding unknown patterns and cluster data in big dataset. Some applications are on computational biology, social network analysis, sentiment analysis, genomics, statistics and so on.¹¹ In reality, some of this activity can be carried out both as supervised and unsupervised algorithms (e.g. sentiment analysis). Lastly, a further type of ML algorithms are reinforcement learning algorithms which have been trained through rewards in case of desired outputs and/or penalties in case of undesired ones. Some of the most compelling usages concern market analysis (stock market in particular) and price fixing.¹²

Among the software agents built on ML and RL algorithms and particularly relevant for the legal discipline are the contract agents: autonomous electronic agents to whom human users delegate cognitive skills instrumental to activities of contract formation, negotiation and conclusion. Often, software agents of these kind are capable of determining much of the contractual content without human supervision until the final agreement is reached (Andrade, et al. 2007; Chopra and White 2010; Dahiyat 2021). Precisely thanks to self-learning technologies, software agents of this type employed in dynamic environments – e.g. the financial market – can infer standard of conducts that do not comply to the user's instructions. Lastly, we are witnessing a progressive use of AI algorithms in the legal professions both for automated litigation resolution systems and for the so called predictive justice. In the second case, AIs support prognostic judgments

¹⁰ This is one the standard – and most famous – definition of machine learning and was provided by the computer scientist Tom Mitchell in (Mitchell 1997).

¹¹ For a rather exhaustive list of the most prominent applications of machine learning algorithms see, among others, (Das, et al. 2015).

¹² An example is (Krasheninnikova, et al. 2019). It is worth noting that the use of these algorithms can also encourage collusive practices, see (Calvano, et al. 2020).

whose accuracy can be improved by the availability of large amounts of data on similar cases. One very controversial application was the COMPAS algorithm, which concerned the risk assessment of recidivism for the purpose of pre-trial detention (Larson, et al. 2016).

The aim of this overview is to help understand both to clarify the notion of ‘artificial intelligence systems’ and which of these computational artefacts are most likely to be of interest for the scope of attributing an autonomous legal personality. More specifically, the common property of the technologies we are concerned with is ability to deliberate according to evaluations that are, to some extent, independent of those of the designer. In a sense, AIs have a degree of epistemic and practical authority over their own behaviour.

Anyway, as already pointed out, not all above mentioned AIs are intended to be the object of specific thoughts on legal personality, but only those that have peculiar features as regards both their functions, their field of use and the societal impact – both risks and opportunities – of their decision-making skills. In general, it is from the combination of two criteria that we can select the AIs candidates to the entitlement of legal personality, and they are the intrinsic risk associated with the conduct of the AIs and the risk associated with the social impact, affecting rights, security and fundamental values.

Some of the cognitive and operational features of AIs will be exposed below and, in the second chapter, I shall also offer a concrete case scenario involving a trading automated system.

3. AI and law

More and more often humans, in different activities — as users of a self-driving vehicles, consumers making a contract with a bot, physicians supported by a diagnostic system — interact with artificial entities that play an active role in such activities. These entities process information and make choices between epistemic and practical alternatives, make changes in the digital or physical world, and engage in communication. Their action appears to be autonomous, in the sense that it is not preprogrammed by a human designer as a specific response to the particular circumstances of the concrete situation. It is up to such systems to process the available information, using learning and inferential algorithms, and to make assessments and decisions accordingly. An additional complexity is owed to the fact that each such

an agent may be part of a dense network of agents, sharing information and making coordinated decisions (e.g., a fleet of autonomous cars).

This raises a set of issues pertaining to legal liabilities, as well as to accountability and moral responsibility. Who is to be asked to explain and justify the behaviour of such a system? Who may be subject to blame or adverse consequences in case such behaviour turns out to be harmful? Who is to be subject to legal penalties, including damages as well as administrative fines or criminal punishment, when such a behaviour violates applicable legal rules. Is it the owner, the user, the programmer, the manager of the development team, the person tasked with testing and debugging? What happens if an action cannot be traced back to any particular human?

As we will see in more detail in the second chapter, one of the available solutions to address these problems is the attribution of legal personality on artificial intelligence systems. Among the motives for granting this status is the fact that many AIs display a kind of *agency*: they are able to perform actions based on reasons and with a variable degree of autonomy.¹³ As Mireille Hildebrandt points out: “the issue of legal personhood is [...] closely tied up with the notion of human agency as conceived in moral philosophy, which refers to the assumption that human beings act on the basis of beliefs and desires, and can give reasons for their actions” (Hildebrandt 2013, 18).

According to the standard conception, the exercise of agency should be explained in terms of the mental attitude to carry out actions, i.e. intentionality (Davidson 1963; Dennett 1987; Bratman 1987; Bishop 1989). On this approach, an automaton that reacts mechanically to its inputs would not be displaying agency, whereas an entity acting by virtue of a causal relationship between its mental states (beliefs, desires, and intentions) and events in the world would. But applying this approach to AI entities is problematic, for while such entities exhibit behaviour that can be understood and foreseen from the intentional stance, we can still ask whether they really possess intentions and other mental states or whether they just appear to in the eyes of an external observer.¹⁴

¹³ Here I make a philosophical use of the notion of agency and not a legal one. In the legal domain agency has a peculiar meaning since it refers to that set of rules governing actions made on behalf of others. This set of rules, by the way, is very suitable to regulate the use of autonomous agents.

¹⁴ The standard concept is characterized by a reductionist approach to the cause-effect relationship between mental states and events: ‘According to an agent-causal

The definition of agency I shall rely on is the one provided by Christian List and Philip Pettit, i.e. the sum of three features of an entity: (1) the ability to represent the environment in which it operates (*representational states*); (2) the ability to know if the external environment should be modified in accordance to the goal pursued (*motivational states*); (3) the ability to implement the action plan elaborated thanks to the elaboration of the previous states and “suitably in the environment whenever that environment fails to match a motivating specification” (*responsive states*) (List and Pettit 2011, 20).

In the next sections I will show how a computer agent can be seen and treated as intentional agents. For the time being, it is our interest to point out that the reason why AIs breaks into the market of legal subjects is due precisely to the exhibition of some kind of agency. As other existing legal entities, they may be considered to be acting for reasons that are the causes of their conduct (Chopra and White 2010, 12).

However, the fact that AI systems possess advanced cognitive capacities may not be a conclusive factor. Certain cognitive capacities — in some regards superior to those of existing AI systems — are in fact also possessed by other entities (e.g., animals), and this does not seem sufficient to justify the conferral of legal personality on such entities, as is also emphasised by the argument from marginal cases. Yet what seem to characterise AIs, to the point of introducing new elements into the debate on legal personality, are the sociotechnical profiles resulting from the deployment of artificial intelligence agents, e.g., the marked unpredictability of their decision-making processes and the impact (both positive and negative) that these processes may have on people’s lives, on society, and on the market; the ability of such systems to communicate and network; the involvement of different human players in the production and implementation of such systems – so called ‘many hands’ problem – each with different potential responsibilities; finally, this multiplies the number of subjects and forums for which potential liabilities arise, creating further uncertainty (so called ‘many-eyes’ problem). The many hands problem, which is not new to accountability studies (Bovens, 2007), not only generates uncertainty in attributing

approach, agency is to be explained in terms of a kind of substance-causation: causation by the agent, construed as a persisting substance. On this view, actions are events, and an event is an action just in case it has the right agent-causal history’ in Stanford Encyclopedia of Philosophy, *Agency*, <https://plato.stanford.edu/entries/agency/#AgeIntAct>, substantive revision in 2019.

responsibilities *ex post facto*, but it also incentivises suboptimal equilibria in which nobody feels obliged to prevent or care about possible negative consequences.

These socio-technical factors create challenges in the management of legal responsibility and accountability profiles requiring effective solutions tailored to the various interests at stake. In other words, the issues these systems give rise to call for a fresh analysis order to determine whether a conferral of legal personality is preferable to other, more limited kinds of empowerment and protection.

What is to be stressed is that alongside human actors there is a growing number of electronic agents capable of predicting and making decisions in a large number of areas (finance, the automotive sector, commerce, epidemiology, public policy, etc.). Some of these systems exploit self-learning functions – i.e. ‘machine learning’ (ML) – which optimize computations thanks to the enormous amount of data with which they are powered. ML has been a great change in the field of AI since it has encouraged to program computing systems through sub-symbolic representation of knowledge (bottom-up approach) as was not previously the case, when it was mainly based on symbolic information processing (top-down approach). In very few words, ML challenges the myth that AIs can only operate mechanically and according to predefined instructions. Some AI systems are based on a cognitive architecture, in which different elements can be understood as beliefs that track the environment, and others as goals pursued and intentions to be implemented. Today’s AIs systems are able to cope with uncertain and dynamic environments, adapting to the lack of information, acquiring new knowledge, and making appropriate (rational?) choices.

While general artificial intelligence, able to cover the large set of cognitive skills possessed by humans is not available, nor will it be in the near future, advanced AI systems are not to be located in a far-fetched future scenario, as they are already part of our societies. According to an estimate by Gartner Analytics, 52 percent of telecommunications companies already use chatbots, and 38 percent of health-care companies currently use expert systems for medical diagnosis.¹⁵ Artificial

¹⁵ “Top Trends on the Gartner Hype Cycle for Artificial Intelligence, 2019” at <https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/>

agents can replace humans in several social-life activities or can be delegated to perform very complex tasks.

The harmful behaviour of AI systems can engender obligations for the human agents involved, such as users, owners, and developers. Artificial agents can also participate in legally relevant interactions, producing effects in the sphere of their owners as well as that of human third parties. Their use can produce legal effects, i.e., contractual rights and obligations, for their users (e.g., vendors in electronic commerce) by virtue of their ability to interact with other subjects (human and electronic) entering into legal transactions. Equally remarkable is the capacity of some AIs to generate creative works - paintings, songs, novels, etc. - prompting the European Parliament to consider whether AI-generated artworks and inventions might be protected by copyright.¹⁶

A recent project by IBM is focused on ‘Automated AI Developments’ (AutoAI) working on the use of AI tools to build and program further AI algorithms: “The AutoAI tool automatically analyzes your data and generates candidate model pipelines that are customized for your predictive modeling problem. These model pipelines are created over time as AutoAI algorithms learn more about your data set and discover data transformations, estimator algorithms, and parameter settings that work best for your problem”.¹⁷ Actually, this is something also Google has been working on in the last few years (AutoML) developing AI systems able to automatically code machine-learning algorithms.¹⁸

3.1. Ceci n'est pas une chose

Whether these technologies are promising or not, the cognitive and operational faculties acquired by AIs, as well as sociotechnical issues raised, call us to question the correctness of their legal qualification as mere ‘things’, and especially whether this should continue to be their legal qualification in the future. The answer to the question of whether some AI systems should acquire the status of legal persons has been at

(Committee on Legal Affairs 2020).

¹⁷ ‘AutoAI: Humans and machines better together’, 18 October 2029, on <https://developer.ibm.com/articles/autoai-humans-and-machines-better-together/>.

¹⁸ Some information about this project can easily be found on <https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html> and <https://cloud.google.com/automl?hl=it>.

the centre of a lively debate. Jurists currently seem to prefer to adapt conventional liability schemes rather than extend legal personality. Although a cautious approach seems reasonable for first-generation artificial agents, this may prove to be insufficient as their autonomy increases (Dahiyat 2021).

What calls for reflection is that, unlike passive artefacts, AIs can be delegated to perform tasks without human review, thanks to skills like: initiating goal-directed behaviour, learning from experience, perceiving the environment and adapting conduct to it, and, if necessary, changing the utility function to maximize the user's targets. What I will argue is that we tend to engage with AIs as intentional agents and to assign cognitive-like states to them, so as to explain and predict their behaviour. It is possible to assume that a computer-based artifact acts on the basis of information about one or more states of the world (epistemic states), which it uses to achieve certain goals (conative states) through a rational and flexible decision-making process. Indeed, AIs generally pursue objectives in a way that is not fixed but changes according to their representation of the current state of the external environment and any change in it. The intentional stance (Dennett 1987) allows us to predict the future behaviour of an AI that is acting teleologically, and it gives us a reliable model for explaining some of its deliberations.

There is no questioning here the ontological status of such intentional states — whether or not they actually correspond to the agents' internal characteristics¹⁹— or the role of consciousness in intentionality. The point is that the conduct of complex AIs can be interpreted as the output of independent intentional states. This may have legal consequences on the way third parties interact with AIs: they can rely on the intentional attitudes directly exhibited by the system regardless of the human controller's upstream intentions. Given this premise, for which it is difficult to equate AIs with other objects, the legal solutions that follow may be of a different nature.

From a purely legal classification, AIs are definitely artefacts, and even if they are very sophisticated, this makes them candidates primarily

¹⁹ Briefly, two different views have solidified around intentionality: on one view, if an action can be explained in terms of cognitive states, that is a sufficient condition for attributing intentionality (*interpretivism*); on a second view, an action can be said to be intentional if, and only if, it is actually *caused* by the entity's internal states and these states are functional organized to produce the 'right event' (*realism*). For an introduction, see (Davies 1998).

for the legal category of ‘things’, since it may seem that artefacts are ultimately other-directed: whatever they are or do is ultimately determined by choices of others, namely, their designers, producers, or users. This status has decisive implications for legal capacity and responsibility, setting aside damages caused by evident programming or manufacturing defects and focusing only on those caused by the proper functioning of AIs, or by the interaction between data, inference models, and human parties. Although the nature of a thing can require a different standard of care, tortious and criminal liability for damages caused by things usually falls on the custodian (who may be the owner or also another controller) or producers (designers-developers).

This approach may lead to inadequate legal conclusions in cases in which harmful consequences are caused by an AI system that was nondefective when put on the market, and that was handled with all due care by its users and guardians. Such a system may have engaged in harmful behaviour — e.g., causing a car crash, a loss in online trading — if it has been tricked by the input data, by the limits of its reasoning and learning capacities, as applied to unexpected circumstances. In fact, just as humans are fallible, so are artificial systems, whose learning is based on statistical models. Thus, a system — like a human expert — may commit mistakes even when it meets all state-of-the-art requirements and is deployed to a task that is appropriate to its capacities.

In these situations, conventional liability models may prove inadequate or conducive to short- and long-term negative effects (more on this later). A possible solution to these issues, may consist in creating an organisation having legal personality — e.g., a limited liability company, a single-member company, or even a memberless LLC (under German law) — which carries out its activity through the AIs. Alternatively, part of the company’s assets can be tied to specific AI-driven businesses (under arts. 2447ff. of the Italian Civil Code). But the aim of setting up a separate patrimony through which creditors can be secured can also be served by conferring a separate legal personality on AIs. This could entitle AIs to hold rights and obligations on their own, enter into contracts by themselves, and produce legal effects on third parties — all this, perhaps, even with complete financial autonomy.

However, against this background, the attempt by the European Parliament to include the ‘electronic personhood’ of robots in the

political debate has not been taken up by the European Commission.²⁰ This proposal struggles to gain traction, since it is believed that some of the risks associated with the advancement of AI can be tackled with the liability rules already in force within our legal systems, or by preventive measures and controls (as in the recent EU proposal for an AI regulation). Moreover, the European Union has no competence to decide what counts as a ‘person’ within national legal systems, as such a decision rests with the Member States.²¹

In this thesis, I will explain why the hypothesis of granting legal independency to AIs still arouses interest both in the theoretical and normative debate and how the conceptual analysis of legal personhood can give us further instruments to make law-policy choices on issues of this kind.

3.2. A risk-based regulatory framework for AI: the proposal of the Artificial Intelligence Act

A regulatory framework on AI recently proposed by the European Commission and currently under discussion is worth mentioning, as it reveals what the attitude of European policy-makers will be towards AIs.²² The regulatory proposal, together with the ‘Coordinated Plan on Artificial Intelligence’ (2021), is part of the wider ‘Artificial Intelligence Act’, with which the Commission intends to lay down harmonised rules on AI. The proposal is intended both to provide developers and programmers with clear requirements for introducing an AI system onto the market – as well as users with guidelines and obligations linked to their use – and to boost public trust in such systems.

²⁰ See the European Parliament’s resolution of 16 February 2017, with recommendations to the Commission on Civil Law Rules on Robotics and the European Commission’s subsequent 25 April 2018 outline ‘Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines’.

²¹ This point has also been made by Thomas Burri: see <https://www.euractiv.com/section/digital/opinion/the-eu-is-right-to-refuse-legal-personality-for-artificial-intelligence/#comment-334495>.

²² European Commission, Brussels, 21.4.2021, Com(2021) 206 Final, 2021/0106(Cod), Proposal For A Regulation Of The European Parliament And Of The Council. Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.

To achieve this, it proposes: a risk level classification of AIs; a list of requirements AI systems for high-risk applications must comply with; specific obligations for users and providers (of high risk AIs); a conformity assessment before the AI system is put into service or placed on the market; a set of rules to be enforced after the system is placed on the market; a governance structure at both European and national level.

One of the most characterising elements of the regulation is the risk-based approach, which revolves around the classification of AIs on three levels of risk: (r1) unacceptable; (r2) high; (r3) minimal. Unacceptable risk is associated with: “All AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned, from social scoring by governments to toys using voice assistance that encourages dangerous behaviour”.²³ Systems that, for example, employ subliminal techniques to subconsciously guide a person's behaviour are therefore banned.

High-risk AIs include both those “[...] intended to be used as safety component of products that are subject to third party ex-ante conformity assessment” and “other stand-alone AI systems with mainly fundamental rights implications [...]”.²⁴ Some application of these systems are then illustrated: biometric identification systems, critical infrastructures (e.g. transport), “educational or vocational training, that may determine the access to education and professional course of someone’s life (e.g. scoring of exams); safety components of products (e.g. AI application in robot-assisted surgery); employment, workers management and access to self-employment (e.g. CV-sorting software for recruitment procedures); essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan); law enforcement that may interfere with people’s fundamental rights (e.g. evaluation of the reliability of evidence); migration, asylum and border

²³ This is from the European Commission website on ‘Shaping Europe’s digital future’: ‘Regulatory framework proposal on artificial intelligence’, link: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

²⁴ European Commission, Brussels, 21.4.2021, Com(2021) 206 Final, 2021/0106(Cod), Proposal For A Regulation Of The European Parliament And Of The Council. Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, p. 13.

control management (e.g. verification of authenticity of travel documents); administration of justice and democratic processes.”²⁵

According to the regulation, manufacturers of these high-risk AI systems will have to fulfil a number of obligations to place them on the market. In particular, they will have to provide adequate risk assessment and mitigation systems; ensure that datasets are constructed to minimise risk and possible discriminatory results; and guarantee the traceability of results, including by providing all necessary information to competent authorities to assess compliance. In addition, the same manufacturers should take care to provide clear and adequate information to the user, including on risk-minimising supervision.

Finally, at the minimal risk level are those systems that do not affect, or at least do not risk compromising, fundamental rights and values. Which are also the most common on the market today: predictive maintenance systems, spam filters and AI for video games.

This regulatory framework is certainly ambitious and could have a terrific impact on AI in Europe, producing interesting economic and legislative consequences. Many of the ideas it introduces will be discussed in this thesis and, in general, a risk-based approach will be privileged to circumscribing the area of AIs that can potentially be considered as legal persons.

4. Artificial minds: intelligence, autonomy and intentionality

In this section the technical peculiarities that make AI technologies relevant to law and, possibly, to legal personality will be explored. It should be pointed out that this analysis is not necessarily intended to endorse a property-based approach to legal personality because, as will be seen, the conferral of this status for AIs may be primarily driven by pragmatic and systemic reasons triggered by a range of sociotechnical factors.

Thus, the aim here is to introduce some preliminary information concerning (cognitive and operational) capacities of AIs, also in function of the more extended discussion on the various reasons behind the demand of legal personality.

²⁵‘Regulatory framework proposal on artificial intelligence’, link: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

More specifically, the following pages will highlight how AIs can be considered as real intentional agents, comparably to other legal subjects, and what kind of implications this may have for the recognition of a legal status other than that of ‘thing’.

The cognitive and operational capacities of AI illustrated below will be found instrumental in understanding where some of the responsibility gaps come from; gaps that, as will be seen in the next chapter, constitute some of the main reasons for considering the hypothesis of attributing legal personality to the AIs. In a way, the three cognitive attributes analysed in the next three sections can be understood as being placed on three levels of abstraction, from the general to the particular.

4.1. Intelligence

What are we referring to when we talk about artificial intelligence? We can both describe a scientific field and a series of technologies somehow derived from it. In section 3, an attempt was made to circumscribe the AI technologies under consideration in this thesis to the so-called artificial intelligence systems (AIs), although this notion still suffers from some vagueness. A more detailed identification of the specific AI systems relevant to our research question will be attempted in next sections, using not only factors relating to the technical components, but also that of the social context in which systems operate.

AI as a branch of knowledge studies the way in which human intelligence, or some of its characteristics, can be reproduced in artificial objects. Even though nowadays these artefacts are mostly computer-based, nothing prevents the same goals from being pursued on hybrid systems in the future. A broad enough, and often cited, definition was proposed by the leading AI scientist John McCarthy, according to whom AI: “is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable” (MacCarthy 2007, 2).

As Pei Wang rightly observed, in this way AI follows two tracks: (1) it proposes a theory of intelligence, deferential to human cognitive

qualities and (2) it studies how to implement that intelligence in technological artifacts (e.g. software) (Wang 2006).

It seems therefore that the development of AI research is closely linked to a more general understanding of the phenomenon of human mind. However, this is not always the case. So far, AI devices have been created even with little knowledge about human intelligence and with even less knowledge about how to replicate it. What seems to matter is rather the way we think our intelligence solves certain problems. The purpose is then to formalise our idea of mental processes into computational ones so that a computer system is able to execute them. But as we still have no evidence that intelligence is just the result of computational abilities, we cannot conclude that we are actually reproducing human intelligence. This is a well-known fact, although there are few interesting arguments in philosophy of mind in support of computational foundations of human cognition (Chalmers 2011).

For the sake of argument, we could say that human intelligence is a very complex faculty in which at least the following skills are involved: computation, judgement, creativity, planning, problem solving and interaction with the environment. But again, intelligence as we use it for artifacts may be considered as a metaphor to describe some similarities between human and computer rationality, more than a proper attribute.

As Pei Wang clearly summarizes (Wang 2008), the similarity between artifacts and humans can be observed from different perspectives and each of them also corresponds to a different way of conceiving AI as a scientific field:

By Structure: from this stance we look at human intelligence as the product of the brain; the more we are able to reproduce the neural substrate the greater the probability of building an intelligent artifact.

By Behaviour: from this angle the similarity is evaluated on the basis of the (observable) behaviour of a system – such as “streams of percepts and actions” (Wang 2008, 3) – rather than on its internal characteristics; in this case both the human and the computer agent are considered as black boxes. The Turing Test was built on this paradigm.

By Capability: here intelligence is taken as the attitude to solve practical problems; according to this view, the capability to solve even

very complex issues prevails over the process - more or less humanoid - by which solutions are reached (Minsky 1986).

By Function: not very different from the previous one, according to this perspective a system that uses typically human cognitive functions to solve problems - searching, learning, computing etc. - is intelligent; the way of organizing information and working with it is therefore inspired by human rationality even without mirroring it (possibly performance is improved).

By Principle: this perspective is more idealistic as it assumes that it is possible to identify the principle or the essence of human reason and then reproduce it in an artifact. A principle that cleverly describes intelligence is that of ‘Bounded Rationality’ according to the definition provided by Herbert Simon: humans act intelligently when they optimize, i.e. they choose the best solution given the available information and computational resources (Simon 1955). It follows that intelligence could be defined as the ‘adaptation with insufficient knowledge and resources’ (Wang, Liu and Dougherty 2018).

I think the last conception also inspire the definition provided by Norvig and Russell of what a *rational* agent is: “For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has”(Russell and Norvig 2016, 37).

Hence, far from anthropomorphic conjecture, the use of the intelligence metaphor evokes one of the most human trait of rationality: the ability to act and decide under scarce information. In fact, the interest in AI is not aroused by the computational power alone but by the possibility that these systems can adapt to incomplete scenarios and act on the basis of plans autonomously developed. Yet, for this autonomy to be a feasible option, AI must get the best out of a flexible type of rationality.

This is a kind of cost-effectiveness principle that inspires *machine learning* (ML) through which scientists provide AI tools with a kind of experience from which they take advantage of the environment and optimize the results – particularly in *reinforcement learning* (RL) – without being explicitly programmed for each single step. Thanks to cloud

computing technologies, then, the experience on which AIs are trained is not necessarily 'their own experience', but that shared by other computer systems (Guoqiang, Wee Peng and Yonggang 2012; Zhihui, et al. 2017). Such interconnection can also take place between single devices that cooperate by exchanging information and data, typically short-range, as is the case with Internet of Things (IoT) technologies.

Thus, even though the creation of human-based artefacts seems postponed due to insufficient information about the nature of human mind, what we call human intelligence is still an object of comparison, as it is expected that by emulating its functions artificial devices can solve specific tasks and complex problems.

Some might argue that the real target of AI research is not much human intelligence, but something *beyond* it. This belief has been raised by the theories of the so-called 'singularity' which predict the advent of an artificial intelligence capable of surpassing human intelligence, perhaps to the point of becoming incomprehensible to us (Vernon 1993; Kurzweil 2005). Speculations of this kind are based on the – debatable – belief that the exponential growth in the computing power of AI technologies may imply a gradual qualitative shift in the type of cognition of AI. However, development does not take place in a generalised way, i.e. it does not affect the general intelligence of AIs, but is rather always directed towards single, specific skills. So, for instance, the improvement of a self-driving car's ability to visually recognise road signs does not cause similar improvements in its understanding of the deontic values behind those signs.

In this context we can be satisfied with a narrow conception of intelligence: intelligence is basically used to signal the ability to act rationally, such as to act according to the goal-directed behaviour and given certain constraints (Nozick 1993); otherwise stated, the behaviour is consistent with the objectives. The broader concept of intelligence can be used as a metaphor – maybe of the *personification* type – and not only in its rhetorical form but as a cognitive tool whose purpose "is understanding and experiencing one kind of thing in terms of another" (Lackoff and Johnson 2003, 9).

Intelligence remains a very general cognitive property, but at a lower level of abstraction we find that of autonomy.

4.2. Autonomy

Is there a reason why qualifying an AIs as ‘mere things’ is more puzzling than with any ordinary (nonintelligent) artifact? The intuitive reply is that an AI makes autonomous decisions, while an ordinary artifact, no matter how complex, mechanically performs what its humans users or designers have predetermined.

Consider the following comparison: a financial agent and a washing machine. Both have been designed to perform functions and achieve goals in a fairly predefined manner. Both execute a significant number of tasks without the intervention by the human agent. The first monitors stock prices and handles one or more stages of the trading process at high frequency, i.e., it places, buy or sell orders when it detects that market conditions are favourable and user-defined requirements are met (Nutti, et al. 2011). The second washes the laundry according to the chosen washing program. Both have been designed to perform functions and achieve goals in a fairly defined manner. Both can execute a significant number of tasks without the intervention (or supervision) of the human agent. Are they both somewhat autonomous? For sure they are both *automatic*. Yet there is something very different between them.

At first, we can frame the differences by reference both to the type of actions delegated and to the way actions are carried out. As a matter of fact, we are inclined to think that the financial bot runs much more complex computing processes than a simple washing machine. While this is intuitively convincing, it is not by itself sufficient to explain the difference between the two artifacts. Indeed, we could easily imagine a much more sophisticated washing machine that, besides doing laundry, selects the best washing setting depending on the clothes to be washed, which is a complex job as it implies the recognition of fabrics, colours etc.

Therefore what counts to distinguish a financial bot from a basic machine is the *way* tasks are carried out. While the second only performs its functions according to the deliberation of its users, the first can make decisions on the basis of further elements, unknown to users, which it acquires from the environment, builds by itself, and considers in its deliberation (market data, predictions based on these data, decisions based on data and predictions) (Tucnik 2010). In short, financial agents can generate deliberative processes that are potentially independent of those of the user on whose behalf they operate. And this happens even when the objectives to be pursued are fixed — e.g., maximising profit

— since the decision path or the course of action to be followed is one the artificial agent computes autonomously (albeit within certain limits). AI trading agents “to play the market effectively, [...] must make decisions in real-time in uncertain, dynamic environments. Successful agents rapidly assimilate information from multiple sources, forecast future events, optimize the allocation of their resources, anticipate strategic interactions, and learn from their experiences” (Chopra and White 2010, 7).

Adaptation to changing market circumstances is now possible thanks to reinforcement learning algorithms and convolutional neural networks (Azhikodan, Bhat and Jadhav 2019). Trading bots can generate deliberative processes potentially independent from those of the user on whose behalf they operate. In short, while some artifacts are autonomous, others are only automated.

Even though an AIs can be considered intelligent according to different criteria, what is most noticeable is the ability to change behaviour according to environmental stimuli (often as an incentive to improve performance). Hence, even if AIs follow set goals they can autonomously pick the most suitable and convenient course of action within the context in which they operate.

There does not seem to be any basis for speculating about the free will of AI, since we know that there are causes that entirely determine the behaviour of these computer entities; the code first of all.²⁶ Moreover, while the variable of consciousness plays a fundamental role in our experience of free will, the same could not be claimed for AIs. So, we can take for granted we are starting from a deterministic viewpoint.

4.2.1. Is there autonomy? A closer look to AI agents

Given that, does it still make sense to believe that AIs act autonomously? Perhaps an answer can be sought by looking at the architecture these systems follow for selecting actions and decisions (Chopra and White 2010, 174). The basic functioning of AIs consist in following a “finite series of well-defined, computer-implementable

²⁶ Someone could use the genetic code to infer the same about the free will of human beings.

instructions to solve a specific set of computable problems".²⁷ This procedure – i.e. an algorithm – allows computer systems to automatically generate predictable output through a given sequence of states.²⁸ In this sense, it could be argued that the artifact does not really operate autonomously as what it does is determined by the coder's knowledge and instructions (*reflex agents*).

Yet, reflex agents are not the only available technique for implementing computer processes. For some classes of problems it is preferable to use flexible algorithms. These problems are inherent to situations that are neither static nor fully observable – e.g. stochastic environments – and where the level of uncertainty is much higher (Russell and Norvig 2016). These are the environments in which, for instance, trading bots operate, since it is very hard to predict what the market trends will be. But also, the environment in which self-driven cars navigate, as no one knows *a priori* what the behaviour of other drivers and agents will be (multi-agent environments).

In these scenarios it is computationally too demanding to lay down all possible sequences of actions and responses to external conditions (*laziness*); moreover, there are objective limits with respect to the knowledge of all possible states, current or future, of an environment (*ignorance*) (Russell and Norvig 2016). This is the reason why operational autonomy of AIs, rather than being a secondary skill, is precisely the added value of these systems. Somehow, this pushes us to accept the risk that a software agents may act unpredictably, since the conditions under which they operate are unpredictable, as well as the specific adaptation that will follow.

All this emphasizes that, even though the behaviour of an artificial agent can be said to be largely predetermined, variables of the environment in which it operates and with which it is called upon to interact, make it impossible to predict the course of actions and deliberations it will take. Residual errors and unpredictable outcomes may also result from incomplete data or a defective infrastructure.

We previously claimed that what makes an artifact autonomous is not much the ability to perform tasks automatically, but rather the

²⁷ Definition of The Definitive Glossary of Higher Mathematical Jargon, entry: Algorithm. Math Vault. <https://mathvault.ca/math-glossary/#algo>.

²⁸ National Institute of Standards and Technology, entry: "Nondeterministic algorithm". <https://xlinux.nist.gov/dads/HTML/deterministicAlgorithm.html>.

adaptation to changing circumstances with lacking information. How does a computer agent who operates under these conditions behave then? It could plan its behaviour either by taking into account all possible states of the world or by picking one or more actions according to desirable situations (*goal-based agents*) and expected utility (*utility-based agents*).

The first option clashes with the so-called ‘qualification problem’: the designer considers all exceptions, even those that are not particularly relevant (Russell and Norvig 2016). This can lead to a combinatorial explosion and, consequently, to immobility. On the contrary, in latter cases the agent computes the probability of possible outcomes and then selects the course of action most likely to achieve the goal or maximize the expected utility (Burr, Cristianini and Ladyman 2018). For instance, if I delegate a bot to buy a used car according to my preferences – e.g. cost, size, type, etc. – then the bot will calculate the utility function that describes my preferences and try to satisfy them with the highest degree of probability. Goal-based and utility-based computer agents display a kind of ‘bounded rationality’: they calculate trade-offs where there are conflicting goals, also assessing the chances of success of one choice over another (Russell and Norvig 2016, 53).

Uncertainty can also be handled thanks to the (self)learning functions with which some AIs are programmed. What makes this learning possible? As briefly mentioned above (section 3), there are several methods: sometimes agents learn from examples or data associated with explicit patterns (supervised learning), sometimes from data with unlabelled patterns (unsupervised learning) and finally through feedback driven by rewards and sanctions (reinforcement learning) (Alpaydin 2014).

Supervised learning is often used for classification algorithms, for example an anti-spam mail filter that, when instructed on the basis of labels assigned to some sample emails, elaborates a general recognition rule that allows to automatically move certain types of emails into spam (Dada, et al. 2019).

Unsupervised learning is much less time-consuming because the coder does not provide examples already labelled, but it is the algorithm that draws unknown patterns from the data. On the other hand, the results are less accurate and the coder often spends a lot of time trying to interpret them. However, unsupervised algorithms are useful to analyse big data trying to find hidden regularities and reduce noise in the data by

combining relevant information (i.e. clustering) or to detect anomalies in dataset (e.g. fraud detection).

Finally, in the case of *reinforcement learning*, the algorithm is not trained on a dataset (although a minimum of background knowledge is often required), but learns from the experience it gets by trying to reach the input target. To this end, the algorithm tests different strategies of action (or mere calculation) to which the programmer associates rewards or punishments. The goal of the algorithm is to maximize the cumulative reward. The reason why reinforcement algorithms are used is that in some environments or practices, e.g. the stock market but also a card game, it is more expensive for the programmer to instruct each single position or behaviour rather than just judging an output produced by the algorithm as positive or negative. The applications of the reinforcement algorithms are numerous, some of the most widespread concern natural language processing, gaming, advertising and dynamic pricing. Of course unsupervised and reinforcement learning are characterized by the greater autonomy with which the agent extracts knowledge from the datasets to achieve optimal behaviour.

To conclude, when we wonder whether an entity is autonomous we are actually asking if its behaviour: (a) is shaped by the entity's experience (b) is self-generated rather than externally imposed and (c) can be modified in virtue of general interests or the peculiarities of the environment in which it acts. Although with different graduations, artificial intelligence systems can virtually meet all these conditions as they act both according to past experiences and to current perceptions of the environment. Their structure can be more or less complex – e.g. model-based, goal-based and utility-based agents – and more or less able deal with lacking information.

As a result, in addition to being rational, can be said to be autonomous insofar as they are able to manipulate, supplement, and expand their initial epistemic and practical knowledge when it turns out to be incomplete or inadequate for the operating context.

Of course, the autonomy of an AIs does not meet the most rigorous requirements associated with human agency – e.g. systems do not set themselves goals, nor do they reflect on the goals assigned to them – but in being capable of self-governance these systems already exhibit innovative elements when it comes to their regulation (Wheeler 2020). And the capacity to engage in reasoning produces a level of autonomy even when targets are set. It is, after all, a matter of degree more than

clear-cut classification: the greater the readjustment capacity in case of scarce information, the greater the independence of the artificial agent from the programmer's instructions and, as a result, the greater the autonomy.

Considering the graduality of these skills, some taxonomies of the levels of automation for automatic systems have been proposed, as we will see in the next section.

4.2.2. Degrees of automation

Since AIs are not all the same, frequently a variety of criteria are adopted to measure their autonomy. The concept of 'Level of Automation' (LoA) was introduced for classifying the various forms of interaction between human operators and machines. Numerous taxonomies have been proposed also taking into account the different areas of use of automated systems (Frohm, et al. 2008). Since our interest is on artificial intelligence systems we shall look at those taxonomies that concern computerised systems and not generic machines.

Automation taxonomies tend to differentiate the level of autonomy not only in absolute terms – where at level zero we find the unassisted human user and at maximum level the machine acting without anyone to drive it – but also in relation to the individual tasks that are performed. One of the most accurate models of this type has been proposed by Parasuraman, Sheridan and Wickens and distinguishes four different steps in the human-machine interaction (Parasuraman, Sheridan and Wickens 2000):

- (1) *Information Acquisition*: sensing and registration of input data;
- (2) *Information Analysis*: working memory and inferential processes for prediction of future trends;
- (3) *Decision and Action Selection*: choice of decisions/actions among alternative and available courses of action;
- (4) *Action Implementation and Adaptation*: execution of concrete actions and adaptation in case of changes in circumstances;

All these operations can be automated with different ranges of independence, e.g. we may have a computer system almost autonomous for what regard the information acquisition and analysis while practically incapable of implementing actions.

This division of tasks into four steps also occurs in another taxonomy suggested by Endsley and Kaber (Endsley and Kraber 1999). In this taxonomy ten levels of autonomy are defined, from the lowest level which corresponds to a purely *manual control* (1) to the level of *full autonomy* (10) in which all functions are carried out by the computer system.

As we may ascertain from this taxonomy, the levels of automation we are most interested in tend to be those from the eighth onwards. It is in fact there that we start to experience a significant involvement of the machine, such as not only in the generation but also in the selection of one of the available decisions.

Taxonomic models about the level of automation have been also elaborated specifically on some AIs, as in the case of self-driving cars. As we see in Figure 1, one of the most widespread versions is that of the Society of Automotive Engineers (SAE) which has defined the following six levels of autonomy for self-driving vehicles. Also in this case, the levels of automation we are concerned about are from the third stage

onwards as it is in those conditions that the human driver is excluded from continuous control over driving.



















SAE Level	Name	Steering, acceleration, deceleration	Monitoring driving environment	Fallback performance of dynamic driving task	System capability (driving modes)
Human monitors environment	0 No automation the full-time performance by the human driver of all aspects of the dynamic driving task, even when enhanced by warning or intervention systems				n/a
	1 Driver assistance the driving mode-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the human driver perform all remaining aspects of the dynamic driving task.				Some driving modes
	2 Partial automation the driving mode-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the human driver perform all remaining aspects of the dynamic driving task				Some driving modes
Car monitors environment	3 Conditional automation the driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task with the expectation that the human driver will respond appropriately to a request to intervene				Some driving modes
	4 High automation the driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task, even if a human driver does not respond appropriately to a request to intervene				Some driving modes
	5 Full automation the full-time performance by an automated driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver				All driving modes

Figure 1. Automated and Autonomous Driving, OECD/ITF, 2015

Even if we never reach the level of full autonomy – perhaps because there are choices that we always want to exercise our discretion over – at the third and fourth level we are already dealing with a peculiar artifact that recognizes the external environment in autonomy and acts accordingly.

What should be stressed is that taxonomies related to the level of automation of AIs can be functional to the design of their legal status. The levels of autonomy describe different ways of allocating tasks between the human agent and the artificial agent and, consequently, which entity is accountable for which actions and how. Therefore if, for instance, a self-driving car with partial automation causes damage following an autonomously calculated trajectory, but for which it requested the human user's feedback, who was aware of the risk, then it makes little sense to speculate on the direct responsibility of the AIs. Just

as, on the other hand, it is appropriate to look at the type of interaction between the human and artificial agent in order to establish whether a wrong decision of the former is affected by a failure of the latter.

Then, for which AI systems does it make sense to talk about legal personality? Questions of this kind can be better addressed once the content of legal personality status and its functional properties have been clarified. For the moment, however, we can certainly point out that what can trigger the conferral of legal personality on AIs is the mechanism whereby the delegation of cognitive tasks to these systems implies loss of control and awareness by the human agent over AIs operations, accompanied by the loss of perception of their riskiness. It follows that an AIs may produce relevant legal effects without the necessary awareness on the part of the holder of the centre of relevant legal imputations. So, what justifies, from a strictly technical point of view, the creation of a new centre of legal imputation is that the human *comes out* of the loop.

The greater the computer's interference in the final stages of the deliberation – i.e. selection and implementation – the greater the autonomous agency of the system. We would not claim the same, in fact, for Rigid Systems where even if the decision is generated by the computer, the selection is under the human operator's competence.

It might also be reasonable to identify a sort of minimum level of autonomy below which there is not sufficient evidence to consider legal states autonomous; or, in any case, to introduce rules that differ from those that apply to other "passive" artefacts. For example, those AI systems that do not have an impact on the generation and selection phases of decisions may not be regarded as satisfying the minimum threshold of autonomy for becoming a legal entity.

Along these lines, a taxonomy that includes the specific percentage of incidence of the autonomous system in the individual functions, as proposed by Parasuraman, could prove useful (Figure 2) (Frohman, et al. 2008).

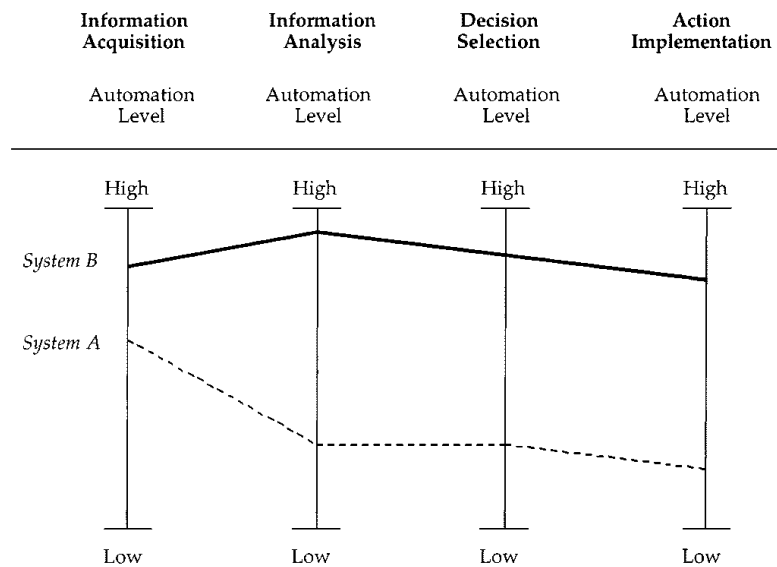


Figure 2. A Model for Types and Levels of Human Interaction

As we notice, although there is interaction between man and machine in both cases, the weight of their roles in the A and B systems is evidently different. Grossly simplified, system B is candidate to become a legal subject.

I believe that taxonomies concerned with levels of autonomy can contribute to the legal reflection concerning agency and personality. However, I would not exclude that alternative ways for classifying AIs (Wooldridge and Jennings 1995), maybe based on parameters like the area of employment or the function type or risk-based, might provide additional criteria for law policy choices of these kind. In some cases they might even make the analysis on the degree of autonomy superfluous, e.g. *filtering agents* that transmit relevant information for their users will likely never become persons in law even when fully autonomous.

4.3. Intentionality: realism and interpretivism

In this section I will claim that from the legal perspective AIs can be treated as intentional agents – or according to patterns of intentionality

– just as other entities qualified as legal subjects are. At the same time I will emphasize that intentionality is not a sufficient condition for legal personality, i.e. it does not by itself justify its conferral on a given entity.

Intentionality is precisely the capacity of minds to be *directed at* or the *be about* objects, properties and state of affairs in the world. In Dennett's words: "something exhibits intentionality if its competence is in some way about something else" (Dennett 2008, 35). The notion of agency I rely on is the one formulated by List and Pettit, for which an agent is "a system with [...] representational states, motivational states, and a capacity to process them and to act on their basis" (List and Pettit 2011, 20). This definition of agency implies that an entity is an agent if and only if it acts intentionally.

But what is it for an entity to have intentional states? At least two different answers can be given.

4.3.1. Realism

According to the realist account, an agent is said to be intentional if it operates according to specific internal states like beliefs, desires, and goals. An agent must then possess an inner cognitive structure that allows these states to be processed. In few words, an entity mental state towards a situation results from the covariance between an internal condition of the entity and the environment, so that the entity can behave accordingly (Dretske 1980). For human and non-human animals this structure, such as the "hardware" that processes the information necessary to have mental states, is obviously chemical-biological in nature, while for artificial intelligence systems may be seen as synthetic.

What is the nature of these mental states is the matter of a classical debate in philosophy of mind that has also benefited from studies on artificial intelligence. One the most prevalent theoretical position is *physicalism*: mental states are to be reduced entirely to physical facts, e.g. an intentional state would be determined by the current configuration of brain cells (and also of other organs) such that this configuration could not correspond to other significantly different mental states (Russell and Norvig 2016, 1028). On this basis, we can follow a so-called *naturalistic* path, that highlights how mental states are the peculiar expression of the biological substratum of the brain, i.e. of neurons, or a *functionalist* one, aiming at identifying mental states with their causal relations and processes (Jackson, Pargetter and Prior 1982).

Of course, these two positions have radically opposite views on the possibility of an artificial system possessing its own intentional states. Since functionalists believe that a reliable model of the mind comes from knowing how the same organises information and carries out operations, nothing prevents a computer program that emulates that organization from being a mind. One of the best known arguments in support of the naturalistic vision is instead elaborated by John Searle with the scope of rejecting the idea that by only reproducing the functional properties it would be possible to create computer minds (J. Searle 1992).

Searle's argument, which challenges the so-called strong-AI scenario, is supported by a thought experiment become popular as the 'Chinese Room' that revolves around the distinction between syntax and semantics. The argument that Searle proposes is that while (human) minds have knowledge of the meaning of symbols, and therefore of semantics, computer programs can only manipulate syntax and emulating the meaning of symbols without really owning it.

The point both types of intentional realists make is the following: intentional states are caused by internal attributes of the entities; the exhibition of intentional behaviour is not sufficient to demonstrate the existence of intentionality, but can at most be a signal of it.

However, this perspective does not prevent us from considering artificial systems as (potentially) intentional agents. Unless we embrace a fully biological-naturalistic view, it is in fact plausible that an AI has internal states of the type of objectives and representations (resulting from computational process of instructions). After all, an AI is autonomous in the sense that it is able to integrate the initial knowledge with new information also sourced from the environment and then implemented.

Also from this perspective, therefore, we can think of an AI as having the following intentional states, and consistently adhering to them:

- (a) **Epistemic State:** information about the world, i.e., about how things are;
- (b) **Conative States:** information about what to do (and how to do it) for changing how things are and then abandon the intentional state once the result has been achieved (Sartor 2009)

(c) **Responsive States:** information about how to process knowledge to implement its goals, even within a changing environment;²⁹

4.3.2. Interpretivism

Some other philosophers believe that the behaviour exhibited by an entity is in itself an evidence of intentional agency. Their core argument is that “all intentional properties are interpretation-dependent” (Kriegel 2010, 113). For this reason, this second account of intentionality goes under the name of *Interpretivism* and is generally associated with authors like Alan Turing, Donald Davidson and Daniel Dennett (Turing 1950; Davidson 1979; Dennett 1987).

Namely, let us take Dennett’s view on intentionality which is framed through the idea of the intentional stance. Dennett distinguishes three different strategies to explain the behaviour of entities and artifacts: the physical, the design and the intentional stance (Dennett 1987).

According to the *physical stance*, we can predict the conduct of certain entities thanks to its physical constitution and the laws of nature governing its existence (physics and chemistry are mainly involved). In this way, we can both explain the behaviour of animals and objects. We can explain why, for instance, the Titanic sank into the ocean after colliding with the iceberg. We cannot explain everything through physical laws since, first, we do not have a global understanding of physical phenomena and, second, some behaviours do not seem to be entirely reducible to them.

According to the *design stance* – or teleological stance – forecasts about the behaviour of an entity can be made assuming that it will attempt to perform functions (and achieve goals) for which it was designed. Of course, this perspective seems particularly appropriate for artifacts, like AIs, that have defined tasks and where each component has a given functional role. Indeed, some empirical evidence suggests that when we interact with artificial intelligence systems we are inclined to prefer the design stance to the intentional one (Marchesi, et al. 2019). So, we can

²⁹ This account looks consistent with Bratman’s model of Belief, Desire, Intention (BDI) (Bratman 1987).

explain actions an autonomous car will take in accordance to the way it has been designed to work without knowing the inner computer processes. And so, for instance, we can predict that the driving speed will tend to decrease whenever the visibility of the road is not at its best.

In reality, the behaviour of biological entities is subject to this stance too. We may actually consider the evolutionary process to be the architect of biological entities' design. Predictions on how a biological organism will behave can be made regardless the awareness bio-chemical aspects: e.g. a virus will try to replicate itself as much as possible on biological hosts.

Social entities such as organisations of people, public or private, can also be interpreted in the light of the design stance. We might explain the behaviour of the Italian Constitutional Court by virtue of the function it performs in the institutional framework and, therefore, foresee that, for instance, it will only deal with issues relating to constitutional profiles. Or, as Giovanni Sartor correctly points out, a commercial company activity can be explained from to the objective of producing profits: “[...] by providing the following reasons: (a) the company has been created by its partners with the purpose of producing profits, and (b) the evolutionary mechanisms of a market economy eliminate companies that do not produce profits and lead to the imitation of most profitable companies” (Sartor 2009, 260).

Finally, according to the *intentional stance* we can make sense of the conduct of an entity by thinking of it as a rational agent equipped with mental states like beliefs, desires, preferences and so on (that attitude of “aboutness” we mentioned).

What distinguishes the intentional stance from realists' idea is that what counts for viewing an entity as intentional is not here the actual existence of specific inner properties, but rather the chance that its behaviour is successfully interpreted *as if* it were intentional. In this sense, intentional states depend on the interpretation of an ideal observer. Extending this position to the extreme and concluding that intentional states do not exist as such, Dennett's interpretivism can be seen as a subset of eliminative materialism.

Dennett suggests that there are two ways in which the interpreter can attribute intentional attitudes to entities: “Here two chief rivals have seemed to emerge: one or another Normative Principle, according to which one should attribute to a creature the propositional attitudes it “ought to have” given its circumstances, and one or another Projective

Principle, according to which one should attribute to a creature the propositional attitudes one supposed one would have oneself in those circumstances” (Dennett 1987, 342). In both cases, the strategy consists in ascribing cognitive states in order to explain and predict conducts of complex entities. Let us consider the example Dennett uses: a computer program playing chess. By putting ourselves from the perspective of the human player challenging the computer opponent, we would not be inclined to use the physical or the design stance to predict its moves and plan ours. Instead, we shall tend to treat him as a rational agent acting with the aim of winning the game (i.e. desires) according to the representations it has of the match conditions and the opponent's moves (i.e. beliefs and responsiveness). So, for instance, we do not infer the specific opening tactic the computer chooses – e.g. the Sicilian Defence – from its design architecture or its physical bases, but rather from the attribution of intentional states, e.g. the intention to respond to the attack protecting the two central pawns.

The intentional stance can certainly coexist with the design and physical strategies and sometimes it looks like a sub-optimal strategy for explaining the behaviour of simpler artifacts like a washing machine. Yet, it has much greater predictive advantages for complex artifacts or entities like human and non-human animals. Typically, we rely on our ability to explain in mentalistic terms the behaviour of other human beings while ignoring the biology of the brain or the function of its parts (even though we might resort to some sort of evolutionary design). What is more, the intentional stance may have an significant explanatory role for group agents. Actually, this strategy likely improves the merely functional interpretation (design stance) of the behaviour, for instance, of the commercial company we have seen before. The design features of the organization can be integrated with the intentional attitudes according to which the group interacts with other market agents and reacts to the environment in which it operates. In case of a market crisis we can better predict the behaviour of a commercial company by considering the aggregate intentions of the group members – the fear of losing its dominant position and leaving a share of the market to its competitors – rather than just the way the same is designed or the objective it purposes.

Also, the intentional stance may not always be totally reliable and this happens whenever empirical generalization and conventions regarding intentional states and their content are betrayed – e.g. irrational

behaviour – which may be caused by failures that need to be addressed through design or physical stances: e.g. a self-driven car that shuts down whenever reverses likely has a fault that should be observed at the level of internal working – design or physical – rather than interpreted as an intention.

In any case, the intentional stance is a coherent way to predict the behaviour of complex artifacts like AIs – unlike the merely passive ones – and most important it also seems to suit the way law works and how it leads us to have certain expectations. Since they have no clue about the algorithm or the computational features, both the owner/user and the counterpart of an AIs will interpret its conduct by attributing intentions that are conventionally associated with behaviours and declarations as displayed by the system. When interacting with a chatbot assistant of an airline company the most effective explanatory strategy is to assume that it rationally takes our questions as the cause of its own beliefs and desires: it desires to give us information to which it has access and it believes that acting in a certain way – e.g. using correct language and data – it will succeed in achieving its objective. It would definitely make sense to describe it as a rational and intentional entity, i.e. acting on the basis of its reasons to achieve its ends (Chopra and White 2010, 164).

An important analogy occurs between group agents and AIs: both act without a conscious mind, but *as if* they had one. Thus, if the lack of intentionality in the realist sense is an obstacle to conferring legal personality on AIs, it should also be an obstacle for collective legal entities; but this does not happen to be the case in most legal systems.

4.3.3 The role of consciousness

For both readings an entity is intentional if it can produce the appropriate rational determinations to achieve certain objectives, except that realists look for the internal conditions grounding the appropriate mental states, while interpretivists are satisfied with external behaviour without the underlying physicalist ontology. But this is not to say that according to the interpretivist reading any entity that exposes exterior behavioral-like actions would be seen successfully as intentional: e.g. environmental entities and events like a tsunami or the eruption of a volcano would not.

Some might argue that the real divergence between the realistic approach and the interpretivist one concerns the role conscious experience has for intentionality. Consciousness may be seen as a mental property too: an entity is conscious if there is *something that is like* to be that entity or, to put it differently, *something it is like* for the entity to be in a certain mental state.³⁰

In this sense, realists would conceive intentionality and consciousness as strictly connected or, even more, as two mutually dependent properties of the mind. Searle takes this position explicitly: “Only a being that could have conscious intentional states could have intentional states at all, and every unconscious intentional state is at least potentially conscious [...] There is a conceptual connection between consciousness and intentionality that has the consequence that a complete theory of intentionality requires an account of consciousness.” (J. Searle 1992, 51). Searle uses the expression “potentially conscious” to mean the possibility that an entity has, at least in principle, the psychological qualities to be conscious.³¹ Searle reaches this conclusion after arguing that intentionality is characterized by having an *aspectual shape* – i.e. intentional states always represent the world under aspects – and that unconscious states don’t have the aspectual attitude that would allow them to function in mental causation or justificatory explanation (Searle 2004, 246).

Yet, it is possible to overcome some this account both through other forms of realism and through hybrid solutions. In the first case it is worth pointing out that there is no two-way correspondence between realism and the “consciousness account”: it is in fact possible to hold a different naturalist theory of intentionality like functionalism – for which intentionality is nothing more than a causal relation between cerebral states and environment states – and its computational variant (Pinker 1997). Either way, what is relevant for intentionality – and maybe also for the supervenience of consciousness – is the functional organisation of neurons. Based on these theories, it is thought possible that artificial systems with certain computational peculiarities may acquire

³⁰ These are standard conception of consciousness in philosophy of mind. They can be traced back to a famous paper by (T. Nagel 1974).

³¹ In modal terms, as Kriegel puts it: “That is, [an entity] M is potentially conscious iff there is a possible world W, such that the laws of psychology in W are the same as in the actual world, and M is conscious in W.” (Kriegel 2003, 275).

consciousness. These positions are often associated with the so-called strong AI trend. However, I cannot go into such theories in depth.

Alternatively, by means of the distinction between *reflexive* and *direct* intentionality we can consider consciousness as a sufficient but not necessary condition of intentionality. Reflexive intentionality would characterize those systems able to thinking about themselves from the intentional stance – i.e. as bearers of beliefs, desires and goals – and consequently whose behaviour can be altered by this understanding (Böök 1999). Direct intentionality would instead correspond to those systems that cannot interpret or think of themselves, but whose behaviour can still be sensibly explained and predicted through the intentional strategy. After all, just like autonomy, intentionality can be conceived in a stepwise manner: low and intermediate levels of intentionality can be conceivable even in non-fully conscious behaviour.

In this way we may keep on considering AIs as intentional agents. The adoption of an intentional strategy for these entities can provoke interesting consequences also in embedding AIs in the legal domain.

4.3.4 Intentionality and legal personhood

Legal personality may consist of powers – e.g. to dispose of rights – and obligations – e.g. to bear responsibilities – which require some form of intentional agency. This is the intuitive link between the two notions, but if we go into further detail we see a dynamic relationship.

At this stage, we know there are two main approaches to intentionality. They differ substantially on the relevance they give to external behaviour: for realists behaviour is only *indicative* of intentional agency, while for interpretivists the former is *constitutive* of the latter. However, they seem to be talking about two slightly different aspects of the same issue and only one is relevant for the legal discourse. In any case these visions are not mutually exclusive.

Even though I tend to agree with realists' view, and therefore to demand a stricter standard of intentionality which is based on the actual possession of the proper mental states, the point is that law does not seem to require the same rigid standard of agency for its subjects.

Intentional stance is, in a certain sense, the strategy that best suits social relations as we often build expectations (and prediction) of the other agents' behaviour on the basis of social conventions. This approach, which applies both to individual agents and to collective ones,

allows us to give a provisional justification of the behaviour of others even lacking deep understanding of the mental disposition of individuals or of each member of a group.

Legal rules reflect the habit of grounding explanation about the agency of social entities on shared expectations; consequently, they seem to endorse the intentional stance. Indeed, both in civil law and, even more so, in criminal law prominence is given to the intentional states of the author of certain conduct (e.g. actions, declarations of will and linguistic acts in general). This is the case of the protection of the counterparties' reliance on the common sense meaning of certain conducts, as it is particularly evident in the rules concerning the formation of contracts (e.g. good faith). Where the intentions of one party do not match those attributed by the counterpart, social habits that best explain the disputed behaviour are assumed to be rules of judgment (Sartor 2009 , 265).

What is to be stressed is that law often does not bother to investigate the cognitive state genuinely formed in the subject's mind, but rather it protects the conventional point of view, even when it may have caused the erroneous belief in other parties (as long as it is reasonable).

Let us now consider the importance that this may have for the conferral of legal personality in general and for AIs in particular. Of course, in the juridical context it would make sense to attribute desires, beliefs and intentions to rational entities capable of acting with purpose. Anyway, being a person in law does not imply being an intentional agent and this is quite evident for many, perhaps for all, legal systems. And the same goes the other way around: meeting intentional agency conditions does not entail legal personality.

Every legal system tends to ascribe personality to subjects who do not have intellectual capacities comparable to those of adults of sound mind, who are one of the most complete expressions of independent legal personhood. Some of those subjects are not intentional at all: e.g. comatose people and unborn children. Hence, a first answer we should give is that legal personality is obviously not just a matter of intentional agency. However, as we discuss below, there are various configurations of legal personality, each embedding different subjective entitlements. These configurations can be grouped in two wide types: an *active* legal personality and a merely *passive* one. Although a more precise picture of such juridical status should be given beforehand, we can now emphasize that intentional agency is a constitutive element of the *active* personality

in law (i.e. holding capacity-responsibility). But not a sufficient one: intentional agents like animals are not (generally) entitled to legal positions.³²

This is not to say that intentional non-human agents have no place in the area of legal personality protection. Actually, it should be mentioned the case of group agents whose behaviour is successfully explained through the intentional stance (as argued in section 4.3.2.) and that are often recognized as legal entities. This is the case for companies, corporations and associations where the attribution of independent personality is permitted by the particular internal organization and the activity carried out. A disorganized community, even if it acts in a more or less coordinated manner, as it is in occasion of a workers' strike, does not create group agency, especially to the legal purposes.³³ On the contrary, a group with an organization functional to the reproduction of intentional states can be considered as a *locus* of agency separate from that of the individuals who compose it (List 2021, 4); which grounds the emergence of an autonomous centre of imputation of legal effects.³⁴ Through the assignment of subjective legal positions the law stabilizes social expectations about collective entities (Teubner 2006). The hypothesis is that AIs display the same new and autonomous *locus* of agency.

Ultimately, the intentionality we have discussed so far is relevant for three reasons. The first is that it is a factor of fraying in the traditional subject-object relationship, which is relevant in terms of the imputation of legal positions and effects. Secondly, it is a determining aspect if artificial agents are to be treated legally as autonomous legal entities (*active* legal personality): as legal rules embed social conventions consistent with the intentional stance, if an artificial agent performs functions according to choices that can be explained and predicted in the light of this strategy, then it can virtually enter into legal relations on a par with other legal entities. Although neither of these reasons implies that intentionality is a sufficient or necessary property to justify legal

³²I am taking as true the account of direct intentionality – i.e. without the constraint of self-reflexivity.

³³ An analogy between group agents and artificial ones is presented by Christian List (List 2021).

³⁴Here grounding is used in its philosophical sense: to indicate non-causal metaphysical relations. This will be deeply investigated in the second part of this thesis.

personality, they play a specific role in the multifactorial evaluations we will see in the next chapter.

In the next section we shall see what other technical peculiarities might make AIs virtually assimilable to other legal entities.

5. The “interaction stance”: AIs’ social skills and legal personification

Rationality and intentional agency are not the only properties that characterize AIs since they also display other skills, namely being part of a network of agents, participating in social interactions and communicating in a comprehensible manner. Actually, this is not new for intentional agency: among the mental states the so called *responsive states* play an important role as regards the capacity to socially interact with other agents within a common environment. This type of interaction can involve both symbolic and non-symbolic communication. As recent experiments suggest it is possible to train AIs to model other agents’ behaviour (robots or humans), in order to predict their future plans of action, through visual processing alone and without any prior symbolic information or even instructions on the relevant inputs (Chen, Vondrick and Lipson 2021). This would open up sophisticated forms of interaction involving correlated strategies and equilibria. If this research hypothesis turns out to be true, it would have a decisive impact not only on the theory of mind for AIs, but also with respect to the degree and type of social interactions that these systems can enact; even the most complex forms of interaction, e.g. competition, which so far seemed to be the exclusive domain of animals with specific attributes, could be genuinely practised by artificial agents.

According to Niklas Luhmann, the ability to take part in social relations is a fundamental element in the definition of agency, as this is ultimately constructed by the surrounding social system more than by intrinsic individual qualities (Luhmann 1996). Under this viewpoint, also legal personality might be the result of the process of attribution of a role by the social-juridical system to single persons or chains of communications (in the case of groups): “[...] it is the market that constructs firms as collectives, otherwise they are nothing but bundles of individual contracts” (Teubner 2006, 501).

Carrying on with a Luhmannian reading, Gunther Teubner addresses the issue of the private law status of electronic agents in the same way as for collective agents. The social substratum of collective legal entities is in fact given by a chain of decisions that communicate both to itself (it “self-describes”) and outside (what Teubner calls “communicative events”) (Teubner 2018). In connection with this, he argues that the legal personification of collective entities was intended to cope with the uncertainty concerning their identity facilitating the interaction with these kinds of social agents; this view looks to me consistent with the intentional stance *à la* Dennett.

The process of personification equates human entities to non-human ones: “Personality means nothing but the symbolic signification of the capacity to participate in communication, and it does not matter and it is historically variable whether the relevant entities are gods, animals, spirits, robots or humans” (Lorentzen 2002, 105). The same would apply, Teubner argues, also to electronic agents – i.e. AIs – as new legal actors emerging from the socio-technical context: “Software agents – just like companies and other formal organizations – are nothing more than mere streams of information that become “persons” (or sub-persons) when they build up a social identity in the communication process and when they are effectively attributed with the ability to act, together with the necessary organizational arrangements, e.g. rules of representation” (Teubner 2006, 120). AIs can be part of social network because they are responsive to communication, as well as being able to represent and aggregate the interests of others or possibly have their own.

I am not sympathetic with this kind of purely constructivist interpretation of agency; what I would like to borrow from it is the instrumental reading of the personification of law.

Interaction skills are relevant to the attribution of legal entitlements for the simple fact that it indicates participation within a network of social actors who recognize each other's rationality and whose conducts may affect the each other's legal sphere, both in terms of benefits and damages (Laukyte 2017). And when non-human entities (as groups) are able to produce effects in the legal sphere of others thanks to a distinctive communicative and decision-making structure, legal systems tend to formalize their internal and external relations through the attribution of an autonomous legal status. This profile is also essential

for setting economic relations and drive transactions between social actors.

To sum up, this should be viewed as an alternative reading which emphasizes how the process of personification also consists in stabilizing social expectations towards certain agents and redistributing the risk caused by their decisions (even if the risk is only causal). Intentionality and the ability to take part in socioeconomic interactions characterize AIs and most of the legal subjects; these affinities suggest that the legal personification of AIs rests on strong cognitive-behavioural attitudes.

6. Conclusions.

In this chapter I addressed some preliminary profiles both with respect to the legal personality and to artificial intelligence systems as potential new legal actors. In particular, I tried to circumscribe the area of AI technologies to those autonomous devices that can actually arouse regulatory and philosophical interest. This task required specifying which technical features make such devices different from ordinary artifacts and how these peculiarities affect their current legal status and their hypothetical one.

Intentional agency and the ability to be part of social interaction are two attributes that are normally correlated with being a legal subject. I formulated this argument in terms of possibility and not of necessity as there is no direct connection between having the above capacities and being an independent legal subject. In any case, it turned out that the relevance of intentionality is not reduced to the investigation of intrinsic or natural qualities, but rather arises in the conventional practice, not unrelated to law, which assumes certain entities' behaviour is best understood as intentional even where there is not – or cannot be – full cognition of its “mechanics”.

While the observations made so far are mostly sociological, rather than juridical in a strict sense, in the next chapter I shall describe the most significant legal implications of artificial systems' agency and the alternative regulatory scenarios to be taken into account.

II. Toys or Subjects: Seeking the Legal Status of AIs

1. Case scenario: Herbert the trading agent

Mrs. Parks is the owner of the 'Four-Wheeler House', one of the major car dealership buying and selling used cars in Florida. The company has an online retail portal to conclude transactions with potential buyers and sellers who interact with Mrs Parks' employees. However, during the weekend the number of customers trying to interact with the 'Four-Wheeler House' grows exponentially to accommodate up to 500 operations per hour. This is not so much a problem for the platform that can safely handle the online traffic, but for the employees who manage each transaction.

In order to efficiently dispose of the huge transaction load that the company receives during the weekend, Mrs. Parks decided to invest in 'Herbert', a trading program made by a software development company and designed to process numerous transactions simultaneously.

Herbert is a trading computer agent that carries out operations with a variable degree of autonomy, by virtue of trading skills knowledge both implemented by coders and independently acquired thanks to internet access and navigation. One of the peculiarities of software agents of this kind is their ability to meaningfully communicate with customers and interact strategically with them. There is no need for human operator to monitor trading agents because they have been trained to understand

communicative stimuli, to look for the response on the internet or in their database and to infer the best outcome.

Ms. Parks has delegated the software development company to program the algorithm according to some rigid guidelines that adhere to the business policy. In particular, Mrs. Parks wants the agent to respect daily budget limits. As for the rest, Herbert can look for relevant information from other sources and also from the market in which it operates. The software also uses flexible decision-making rules thanks to machine learning algorithms that train the agent on a dataset consisting of the Four-Wheeler House's past transactions and of its own stored experience.

Thus, the trading agent can conclude transactions even in the absence of explicit coder (or user) instructions; rarely each operation is solvable through static rules and, by the way, it is Herbert's peculiarity to decide with partial autonomy. More specifically, Herbert works in this way: every weekend Mrs. Parks determines what overall business objective it must pursue, i.e. sometimes it may have to sell more cars than it buys, other times it will have to focus on the bargains and so on. Given the initial goal, the software will have to evaluate the price of each car – either to be bought and sold – using and combining some parameters: list price, brand, mileage, year of registration, engine type and emission standard etc. Besides these fixed parameters Herbert also provides market information it finds on its own on the internet: for instance it can find out that for a specific car model there is a shortage of spare parts at the suppliers, which negatively affect the final price. Once the price is set, Herbert calculate and predict which trading strategy maximizes the overall business objective also weighing different alternatives: e.g. if the objective is to maximize profits without taking up space in the car depository, the software will only buy those vehicles that it believes it can immediately resell at an additional low price.

The last step is that of the actual purchase or sale offer and the ensuing negotiation with the client. Herbert conducts the negotiation without human review, which means that in case of a counteroffer it will evaluate autonomously – going through all the decisional steps – whether to accept, reject or reply with a new offer. If everything goes well, the negotiation ends with an agreement that will be transmitted to human operators because it lacks the legal capacity to conclude a contract.

For both sales and purchases Herbert has learned several commercial strategies and occasionally manages to predict the most suitable one for each transaction, also by exploiting the counterparty's personal information (lawfully). In particular, when operating as a seller, Herbert learned a business strategy that proved to be particularly successful: it puts an extremely under-priced car, A, on sale carrying out multiple transactions at once on the same; however, after selling the discounted car A to the fastest buyer the algorithm has learned that, instead of giving immediate notice of the purchase to other customers in negotiation, was profitable to wait to do it just before the (potential) final agreement. At that point the algorithm notifies that the car has already been sold but instead of withdrawing from the sale it offers an older model, B, at a price even lower than that of A. This is because Herbert has developed an expected utility, i.e. choices between probability over outcomes, on this latter model of conduct: the probability of reaching an agreement increases by 30% compared to proposing B directly.

Unfortunately, Herbert's behaviour amounts to an unfair commercial practice according to the Directive 2005/29/EC of the European Parliament.³⁵ The software agent has escaped the constraints of a not impeccable programming but in a rather unpredictable way – both to the coders and to the user – and has taken advantage of the temporal lag that intervenes between the conclusion of the sale and the notification to the other clients in course of negotiation. In short, the AIs has *intentionally* adopted a deceptive revenue-maximizing practice by counting on *bait advertising*.

2. The use of AIs from the legal perspective: risks, damages and responsibilities

There is no such a trading agent on the market named Herbert that behaves in the way described in the previous case scenario. Still, I tried

³⁵ This practice is generally named *bait advertising*. Specifically, point 5 Annex I on *Commercial Practices Which Are In All Circumstances Considered Unfair*: “Making an invitation to purchase products at a specified price without disclosing the existence of any reasonable grounds the trader may have for believing that he will not be able to offer for supply or to procure another trader to supply, those products or equivalent products at that price for a period that is, and in quantities that are, reasonable having regard to the product, the scale of advertising of the product and the price offered (bait advertising).”

to picture an artificial agent that was, although not in use, absolutely within the reach of current technologies. The unlawful consequences of self-learning skills, as imagined above, are thus not bounded to the fictional scenario. Indeed, there are further case studies that are equally emblematic like damages produced by self-driving cars and high-frequency financial agents.

In this chapter I shall attempt to highlight the regulatory legal aspects involved by the existence of artificial agents capable of acting with autonomy, intentionality and sometimes with unpredictable outcomes. I shall then introduce the debate on the assignment of independent legal personality on AIs.

As described for the previous case scenario, a software agent that is able to acquire information independently and learn (or modify) its utility function from the data it is trained on, as well as from its own experience, can produce unexpected outcomes and possibly unlawful conduct. The harmful outcomes worth focusing on are not those caused by clear negligence on the part of the human agent, but those that emerge from the autonomous agency of the AIs. That is, those outcomes for which the predictive capabilities of the accountable subjects – e.g. programmer, producer or user - are extremely low. It goes without saying that such profiles have relevance for the legal status of these peculiar artifacts, for the way they can be delegated to perform tasks with high involvement of cognitive skills and for the legal responsibility thus following.

Specifically, I will address the issue of liability regimes by which we might govern damages, torts, breaches of contracts and eventually crimes produced by AIs on behalf of (one or more) human users. For criminal law I will discuss scenarios in which AIs commit crimes (with and without personality). For what concerns civil law, I shall present three potential juridical status for AIs which imply different consequences over the allocation of legal responsibility: (a) AIs as cognitive tools (or proxies) (b) AIs as part of a collective legal entities and (c) AIs as independent legal subjects. Each legal qualification distributes civil and criminal liability differently on the agents that gravitate around each AIs: i.e. producers, coders, designers, data developer, operators, owners and users. By the way, civil liability - both contractual and tortious - will be explored in more detail since, as I will argue, for the criminal liability of AIs the regulatory policy appears even

more demanding and the appropriate technological standard still uncertain.

At this stage, one might ask why an analysis of legal personality is so much deferential to liability issues. The answer is that I take as relevant to the research question a *functionalist* approach to personality, i.e. identifying personality with the function it performs and the coordination problems it solves. In other words, what counts are the concrete effects produced, which, depending on the situation and the nature of the status, range from access to fundamental rights to the regulation of sets of legal relations, to cooperation between agents and more.

The distinctive function of legal personhood is the imputation of legal positions, with the associated allocation of benefits and costs that this imputation implies. As is the case of collective agents, i.e. juridical persons, legal personification may be primarily instrumental in coordinating meta-individual activities also through the efficient distribution of related risks and rewards. Liability is the legal position that typically sets out how the costs of harmful conduct are distributed. This holds particularly true for the civil law area. I will elaborate on this point by discussing the functions pursued through the legal personification of collective entities (sec. 5).

Regarding the more general discussion on the ontology of legal personality, I shall not reject the *essentialist* argument out of hand, but this is a tricky point which requires a better understanding of the role of the natural substratum of personhood, that will be addressed later. I will not rule out the possibility that there may be additional reasons that do not fit this logic: e.g. the protection of personal integrity. However, current AI models are not yet at the point of calling for a constitutionally protected personality. So far, instead, the pressure to examine this regulatory option stems from the issues of attribution of responsibilities. Therefore, I shall take the functional role of legal personality as a sufficient justification (albeit hypothetical) for determining whether or not AIs should be qualified as legal subjects.

Yet there are also other ways, besides personality, in which different mechanisms of (civil) liability distribute risk. For example, models based on negligence shift the risk to parties who failed to comply with certain standards of conduct, while market models shift the risk to parties able to bear the cost at the lowest price and so on. Before engaging in the

analysis of specific liability regimes, though, I would like to explain how troubles of responsibility allocation may arise from AIs' activity.

AI failures can have different sources: I would divide them in (f1) mistakes, (f2) misuses and (f3) accidents. The first source (f1) cover discoverable and verifiable defects of the product and errors caused by programmers – both for code design and data training – or by manufacturers; possibly also errors in the design of the infrastructure that supports or in which the AIs operate. The second source (f2) collect misuse practices both on the programming/production side, e.g. lack of information about the usage and performance limits of AIs, and on the user/operator side, e.g. unlawful use.

Finally, I would use accidents (f3) to include both cases in which the failure is unlikely to be traced to a single party – due to complexity, dynamicity and networking of devices, human parties or data, e.g. by virtue of IoT, cloud computing or distributed ledger technologies (DLT) – and cases where failure is highly unpredictable – due to inner opacity and/or self-learning skills (especially if unsupervised); here again, the interaction with the environment, data and inference models may be crucial factors: “There are various reasons why the resulting algorithm may behave in a somewhat unpredictable manner later. Errors in labelling apart, a situation may be characterised by features not represented in the training data (eg vehicle software has been trained with numerous images of oncoming traffic, but there was no image of oncoming traffic against the backdrop of an extraordinarily blazing sundown)” (Wendehorst 2020).

Not wrongly, some suggest that the figures of failure – or at least risk – of AI should also include *structural effects*, i.e. those indirect negative consequences produced by the way in which AI transforms the environment in which it is developed; one scenario to consider is that in which AI could indirectly increase the risk of nuclear war (Zwetsloot and Dafoe 2019). However, although of deep interest, the structural perspective does not seem relevant to our discussion.

This third class of failures (f3) is the most problematic one as far as the distribution of risks and costs is concerned and, given the functional role recognized to personality, it stimulates a broader reasoning on the legal status of these artefacts. It is needless to point out that such failures are more likely among that class of AIs that the Commission has recently defined as high risk (see sec. 4.2, Ch. 1). In this sense one can benefit from the theoretical overlap between the two sets, (f3) and (r2), which

is not total since high-risk systems due to their impact on fundamental values and rights may be much safer than others.

2.1. Accidents will happen

In the previous chapter it was already stressed how the technical macro-characteristic of operational autonomy – to which we can associate the philosophical concept of agency – has a significant impact on law. And we need not to take fully autonomous systems – e.g. level 6 in 5.2.1. – for figuring this out. Autonomous decision-making is mostly the result of the algorithms' training with machine learning methods – in its various forms – and not of the mere computational power. This suggests that the legislators can focus on ML technologies to assess the novelty dimension of AI compared to other artifacts. Of course, problems of this nature are not unique to ML-trained AI systems, but in general to any AIs pursuing goals as also Stephen Omohundro argues (Omohundro 2008). What self-learning mechanisms bring is actually a marked unpredictability of the courses of action that are chosen by the system, especially for the more opaque learning methods, e.g. deep neural networks (DNNs).

Omohundro describes few general patterns that characterize goal-seeking systems and that may contribute to the occurrence of damage. These *basic drives* make AIs to some extent “unreliable” by nature. First, AIs have persistent self-improvement incentives to optimize the achievement of expected goals. The more a system is based on erroneous or uncertain models, the greater the drive to seek benefits from enhanced performance. Yet, editing the software can have detrimental effects on the system, the users or third parties. Indeed, as described in Herbert's case scenario, the tendency to maximize gain creates an unfair commercial practice. This drive for self-improvement is very difficult to limit and counteracting it is likely not the best way to manage the employment of AIs (Omohundro 2008, 485). Secondly, AIs “want” to be rational: they represent their goals according to utility functions – when outcomes are reasonably predictable – or to expected utility – when uncertainty is higher – and tend to avoid becoming irrational; in this sense AIs are rational according to the model of rational economic agents. Thus, if successful, each AIs have an interest in preserving its utility function – without counterfeiting it - as well as the ultimate goals, except in a limited number of cases where change is

expressly contemplated (e.g. when they create proxy systems with different utility functions).³⁶ Third, AIs have a big pressure to acquire as many resources as possible to efficiently accomplish their task. Resources generally means data and pieces of information. As observed for the case scenario, resources can be acquired through exploration and targeted search on the web or through market trading. However, Omohundro points out, resource acquisition is not always done in a positive manner: “Unfortunately the pressure to acquire resources does not take account of the negative externalities imposed on others. Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources”.

These technical attitudes show that the risk arising from the use of AIs, as they are currently designed (and are expected to be designed in the near future), is inherent in the way they work rather than caused by the negligence of the programmers. Of course some unfortunate consequences of AIs’ performance can be avoided with better standards of reliability, but programmers cannot be expected to provide instructions for every single behaviour under all circumstances so that every outcome is determinate; unless by giving up on the very idea of intelligent agents or dramatically reduce the degree of autonomy. Furthermore, for contemporary self-learning AI algorithms that, as anticipated, add opacity and unpredictability to the list of technical peculiarities, the claim of full reliability becomes even weaker as the critical part is not the source code, but rather the dataset. While these algorithms might have transparent codes, absolute control over the data on which they are trained and work is often time-consuming and unnecessary. It is not essential to know, for instance, the exact reason why a computer vision system identifies a siberian husky with a guardrail. What seems relevant is the human ability to detect the error in time to prevent it from recurring. If the failure has already caused damage, then an adequate level of transparency may help to deal with simple cases, while at other times it would still be insufficient for identifying where the responsibility for the malfunction lies. Yet, this should not suggest that the way to overcome problems of this nature is to seek a standard of *full* transparency, as I will argue in the next paragraph.

³⁶ According to Omohundro, these systems also use strategies to safeguard themselves from destruction, e.g., replication or displacement, (Omohundro 2008, 487).

Nonetheless, I believe we should not renounce to the idea of developing autonomous artificial agents and that the risk of damages produced by unpredictable behaviours can be accepted for essentially two reasons: (a) even assuming that perfectly predictable and trustworthy intelligent agents can be developed, there are trade-offs to take into account: e.g. the cost of producing fully transparent AIs is likely to decrease the performance of the algorithms, so the expected gain, which is a disincentive to production and innovation; (b) despite operational uncertainty, AIs typically display a far greater capacity for decision-making than humans, and this is an advantage we should not dispense with (above a given safety threshold). Take as an example an AIs supporting medical diagnosis. There is a utilitarian rationale for accepting risks that are counterbalanced by opportunities to improve the general welfare: as long as the marginal benefit, i.e. better accuracy, outweighs the marginal cost, then there will be a solid justification for adopting these technologies; even for models that are not particularly transparent (i.e. not explainable) (Hacker, et al. 2020).

Thus, we have preliminary motives to consider the use, perhaps provisional, of AIs even under suboptimal conditions of predictability and to design the best allocation of risks and costs of damages. Yet, the objective of this chapter is not to propose original regulatory policies but rather to consider the pros and cons of the available legal solutions – mainly in the civil law area - for the distribution of responsibilities and to highlight the need for reflecting further about the legal personality option. I will therefore hold a *weak* thesis about this last possibility, as summarized by the following points:

1. Although AIs have better decision-making skills than humans, they may produce significant damages by affecting fundamental aspects of people's lives.
2. Harmful conduct of AIs can be unpredictable, untraceable or unexplainable, and thus create accountability gaps.
3. Often there are no short term technical solutions to accountability gaps, but there are good reasons to bear the (social) risk of AI systems making unexpected and untraceable failures.

4. Accountability gaps can be tackled either with or without the attribution of autonomous legal personality on the AIs, but in the latter case with inefficient solutions for all those situations where failure emanates from the autonomous agency of AIs.

5. Conferring legal personality on AIs, under given circumstances, leads to a more efficient allocation of social risks and costs; also simplifying in the imputation of legal effects.

My overall argument is therefore based on the assumption that if independent legal personality is a more efficient and equitable way to balance risks (benefits) and costs stemming from AIs' autonomous conduct, then we should prefer it to the alternative juridical strategies. In the section concerning the reasons for granting legal personality on AIs (Sec. 5), I will provide a more detailed analysis for this argument. What the relationship is to the ontological background will be explored in the second part of the thesis, which is devoted to the philosophical investigation.

3. Moral responsibility, liability and accountability

When we speak of responsibility we can do so with at least two meanings which, albeit often interfering, may have different forms and implications: moral and legal responsibility (also liability here). Moral responsibility in its broadest sense is a system of attribution of rewards or punishments – both ideological and concrete – for the occurrence of certain events. Liability is a system of punishments which arises from a number of situations considered relevant in a legal system – e.g. a contract, a crime, a tort, a non-pecuniary relationship – and results in an obligation to give or do something. Of course, there are differences between liability under civil law and liability under criminal law, as will be discussed below.

Although there is some degree of interference between moral responsibility and liability, the two concepts do not overlap completely. Just think of those cases that in the Italian legal system are known as 'impossible crimes' (art. 49, co. 2 Criminal Code), whenever the unsuitability of the action or the inexistence of the object make it

impossible to damage other people's legal goods: if someone shoots intentionally an already dead body, the offence is impossible and therefore the shooter will not be convicted of murder (or attempted murder), but at most security measures may be imposed on her if the judge considers that she is still a dangerous person. In both cases, even if the Italian legislator does not hold those who carried out the conduct liable (by virtue of the principle of offensiveness), we would still tend to blame them morally because those persons had knowingly decided to cause the damage.

Claiming that a person is morally responsible is to evaluate – i.e. to blame or praise – that person's powers of judgement in the way they have been exercised in a given behaviour or role. Also independently of any causal link with the relevant event. This calls for the entity, whose moral responsibility is preached, to be able to give normative significance to the situations and choices to be made, as well as being able to control them in practice. For this reason, moral competence is often regarded as a precondition for moral responsibility (Wallace 1994). Others make moral responsibility depend on the concept of moral personality (Gordon 2021).

In the case of liability, things are quite different. Liability is often attributed on the basis of criteria that disregard the existence moral personhood and that rather respond to criteria of social efficiency (especially in civil liability). Also for liability, these criteria may be independent of the specific causal link with the event for which liability is sought. Sometimes, the subject is liable purely because it has a duty, perhaps to control or oversight, over a series of processes of which it is not the causal origin. Just think of the case of the editor of a newspaper who is liable for supervising the content of articles published by his employees. Occasionally the connection between liability and causation or fault breaks down altogether, as is the case whenever a party has an obligation, maybe to pay damages, as a mere result of certain legal relations (e.g. strict liability or no-fault liability).

Anyway, both moral responsibility and liability are related to the concept of *accountability* which has been developed in ethics and governance studies – as well as being particularly recurrent in artificial intelligence debates – as an operational notion (Kroll 2020). If responsibility, to a first approximation, results from the connection between actions, faults or duties and consequences, accountability is the requirement to provide information about those and to keep track of

them. Or at least this is the most specific meaning we can attribute to it, since at higher levels of abstraction accountability can appear as a simple fidelity to ethical and legal standards.

Hence, accountability is the requirement to provide information about certain activities and to keep track of them, and can serve as a tool for determining who is to blame for violations of moral and legal standards. More specifically, accountability refers to that specific relationship in which one entity (e.g. agent) is called upon to track and answer for its behaviour – and possible failures – to another entity (e.g. the principal). The latter must evaluate – by approving or disapproving – the merits of the information received. A government, for example, can be said to be accountable to other institutions and to the electorate. Against this background, we often find the concept of accountability identified with that of answerability.

An appealing version of the argument that accountability is an operational feature of moral responsibility – as an instance of explanation exercised with authority – finds expression in Stephen Darwall's theory of the "second-person standpoint".

The notion of accountability can be profitable for framing responsibility for failures in the conduct of AIs. In fact, it seems appropriate for contexts in which a plurality of subjects interact in a complex way as it can help clarify the purpose of the analysis of responsibility by highlighting salient relationships and queries. As Joshua Kroll puts it: "Successfully demanding accountability around an entity, person, system, or artifact requires establishing both ends of this relationship: who or what answers to whom or to what?" (Kroll 2020, 183).

Yet, the accountability relationship can take on very different configurations due to the role played, especially in complex systems, by a large number of elements. This is striking for human-AI interaction and digital services, where the complex inter-play between actors, practices, data and inferential models often makes it difficult to define the contours of accountability. The attention for an accountability analysis can in fact be directed to different fragments of the relevant interaction, especially if we talk about AIs: the organisational context, the ethical or policy guidelines implemented in the tool, the algorithm, training data, inference model and so on.

I would like to stress that as AIs are sociotechnical systems, this affects the way accountability for their conduct is built. The notion of

sociotechnical systems has been developed during the 1960s by the Tavistock Institute for work organisation studies (e.g., factory work), and refers to complex hybrid systems in which human and technical resources are joined in goal-directed behaviour (Baxter and Sommerville, 2011; Long, 2013). From a sociotechnical perspective, an outcome reached in such a system is the result of the interaction between social, organisational and technological factors.³⁷ Hence, the performance of a sociotechnical system relies then on the joint optimisation of tools, machinery, infrastructure and technology (e.g., software), on the technical side, and of rules, procedures, metrics, roles, expectations, and coordination mechanisms, on the social side. The dense interplay between these components prevents their disentanglement as single observable parts in concrete circumstances; just as it prevents the detection of a general function, as such hybrid systems are embedded in a network of individual actions. This distinguishes sociotechnical systems from traditional technological artefacts or mere social systems (Vermaas et al., 2011).

As sociotechnical systems, AIs embody both the technological component and the social one, engaging epistemic and normative assumptions. Modelling the overall behaviour of AIs as the expression of just one of the two components – typically the technological one – leads to ineffective accountability schemes. Indeed, social assumptions can easily be the matter of contestations to which the agent must answer with explanations and justifications, according to the accountability mechanism (Binns, 2018).

Said that, what changes when looking at moral or legal responsibility through the kaleidoscope of (sociotechnical) accountability are the conditions of blame and the relationships with the relevant facts. Putting accountability into practice means precisely making these links explicit and formalising them (more on this in the next section) (Kroll 2020, 185).

We can partially dispense the analysis of moral responsibility both because for the types of events we are interested in – the so-called accidents – the attribution of moral responsibility to human subjects seems to encounter important barriers and because, subsidiarily, directly

³⁷ One way to see the difference between these components is that while the technical components are governed by natural laws, these are not sufficient to explain the social components (Vermaas et al., 2011).

morally blaming AIs would require a moral competence that this technology does not yet satisfy.

The investigation from the perspective of liability is no less intricate. Two aspects should be addressed: (a) on the basis of what criteria a human agent can be held liable for the conduct of an AIs (b) in what sense is it possible to hold the AIs liable in a direct way. The second issue stems from scepticism about the ability to respond effectively to the first point. Note that liability for the fault of others is not new to legal doctrine: e.g. consider the case of the servant in Roman law or the liability of the principal for acts committed by the agent.

Let's stick to point (a) for now. As mentioned in the previous section, we are interested in a particular set of harmful conduct, what have been called accidents. For these situations some of the criteria usually adopted to establish legal liability hardly seem applicable.

We cannot rely on a criterion of causation: it is difficult – sometimes impossible – to establish who caused the damage in a procedure as complex and opaque as that which characterises the decision-making of modern AIs.³⁸ At times, this is due to the so-called 'many hands' – or 'many eyes' – problem and at times to the technical inscrutability of the software (e.g. DNNs).

Nor can we rely entirely on fault criteria, e.g. negligence, due to the impossibility of tracing them or the lack of elements to determine the correct standard of conduct. We will elaborate on all these issues when discussing the civil law status of AI in the next chapter.

In any case, it is needed that accountability functions to a minimal degree, i.e., to be able at least to distinguish situations of error, misuse or accident. What I want to stress is that, although in AI services there are major obstacles to the proper and effective functioning of accountability for fault-based (or causation-based) liability schemes, a minimal reconstruction of this relationship is necessary even for unconventional solutions such as strict liability or legal personality. Failure to do so would provide dangerous incentives and create a malevolent legal shield for any kind of activity.

³⁸ But it is worth reiterating that the case studies are those in which there is no evidence that the action has been caused by the mistake (or the clear intention) of a human being.

4. How to enforce accountability for AIs: transparency and explainability (XAI)

Before considering whether accountability is a useful notion for addressing responsibility issues about AIs activities, it should be ascertained to what extent it can actually be enforced. As was pointed out in the previous section, accountability is a subset of overall responsibility and can serve as a criterion for determining who is to blame for violations of moral, social and legal standards. More specifically, accountability refers to that specific relationship in which one entity (e.g. agent) is called upon to track and answer for its behaviour – and possible failures – to another entity (e.g. principal). In order for this mechanism to function optimally, i.e. to enable the required oversight, the accountable entity must transparently disclose its activities, its organisation and its decision-making.

As a consequence, transparency might be conceived as a precondition for accountability. In complex entities, both social and artificial ones, the more information is available, the easier it will be to identify the source and the reason for a mistake:³⁹ e.g. the more we know about Herbert's software or about the internal organisation of a public administration, the greater the precision with which it will be possible to identify the responsible parties and their contribution. The total absence of transparency makes accountability unfeasible: if there is a lack of information about the decisions and procedures of a stock company, there will be obvious disadvantages in calling potential perpetrators to account.

Transparency could play a key role for artificial intelligence systems because it would facilitate the enforcement of accountability and, as a result, liability. The application of accountability, however, may be uncertain in the field of AI since it may not be easy to establish which subject has the duty to keep track of and possibly intervene in the case of malfunctions. So the first aspect to be considered relates to the selection of the relevant agents as well as the relevant information. In other words, how to identify which actors and information have value

³⁹ Put in this way it would seem that transparency and explanation are equivalent. In reality this is not the case, as I will argue later. There is, however, undoubtedly a relationship between the two. To some extent, one can say that the chance of successful explanation is not invariant to the level of transparency of a system.

for the purposes of attributing responsibility? As Joshua Kroll suggests: “The best way to determine which records best support accountability is to determine what oversight is necessary and to determine how to facilitate that oversight” (Kroll 2020, 188).

It should now be pointed out that transparency can operate both at the level of the *outcomes* that a system generates and at the level of the *procedure* that has been followed (Diakopoulos 2020). Depending on the recipients and the context of use of the AIs, interest may be directed at one or the other level; and the information made accessible is likely to vary.

Either way, monitoring the performance of an AIs requires transparency of organizational – e.g. expected results – and computer elements – e.g. the computational model, training data and the infrastructure in which the system operates. The exposure of both elements would allow human involvement to be traced and located, which would be crucial to weigh the different roles of designers, engineers, programmers, producers, owners and users. The ways human agents are involved in the activities of an AIs range from general supervision, feedback to improve performance, explicit instructions, to direct interventions that break up the autonomous activity (Diakopoulos 2020, 201). It should be recalled that a clear picture of human involvement is crucial for the purpose of legal liability: if we find, for instance, that training data of an algorithm for predicting loan credibility is biased, we will conclude that the users of the system would probably not be able to supervise or correct some of the occurred mistakes. Figuring out where accountability is to be placed is made easier – or even possible – by the fact that the accuracy and neutrality of the dataset can be reviewed. Anyhow, none of this is of benefit if transparency does not entail the AI system's *interpretability*, i.e. if it does not allow a human being to understand the cause-and-effect relationships within it and to predict possible variations when the input data change.⁴⁰ This is a higher standard than simple *understandability*, which is usually conceived as the

⁴⁰ This conception of interpretability is consistent with the definition of algorithmic transparency: “a linear model is deemed transparent because its error surface can be understood and reasoned about, allowing the user to understand how the model will act in every situation it may face” in (Arrieta, et al. 2020). On algorithmic transparency see also (Datta, Sen and Zick 2016) .

possibility to understand the functioning of an AIs without knowing its model, data or internal structure.⁴¹

However, it is not without controversy whether transparency succeeds in its promise of improving the accountability of a system. In this sense, I believe there are two different sets of problems: the first contains problems of concrete implementation of transparent systems; the second one concerns the actual fitness of transparency for assessing legal liability. In theory, in a fully transparent system it should always be possible to monitor decision-making and activity-generating processes and then eventually to discover the reasons behind any accident. Unfortunately in practice it is very difficult, if not impossible, to design fully transparent systems.⁴² What are then the main reasons behind this issue?

The first reason has already been already mentioned in the previous section and is essentially a quantitative problem: some of the most sophisticated AI systems run on the basis of millions of parameters and data (think only of ANNs). It is easy to understand that disclosing such a large amount of data is not a guarantee of transparency. On the contrary, this makes the objective of full transparency, if not unrealistic, at least unnecessarily demanding; it follows that the likelihood of transparency turning into *interpretability* decreases. In other words, even if all parameters and data could be disclosed, the expected benefit would be comparatively negligible.

Secondly, AI algorithms are subject to continuous transformations that can quickly change whole pieces of code and even final outcomes (Diakopoulos 2020, 208). Basically, just as smartphones receive updates that can slows it down or improves their performance, an algorithm can change the dataset and thus the experience it is trained on. Naturally, the dynamic character of these algorithms risks compromising the effectiveness of full transparency, which would require constant monitoring of the time span relevant to the state of the algorithm in use. If the transformations in question occur at the level of the individual artificial system, then the whole enterprise risks being undermined.

Thirdly, transparency comes at a cost both in strictly economic terms and also in terms of performance. This means that a company seeking

⁴¹ For an exhaustive taxonomy of the most recurrent concepts in the literature on explanation and artificial intelligence see (Arrieta, et al. 2020).

⁴² Of course there are fully transparent systems, but they are very simple and have few variables or patterns (often in the form of decision trees).

to develop an AI model will have to bear in mind that there are costs involved in preparing the legal documentation and regularly updating it to ensure that the model is transparent. On the performance side, instead, a trade-off between transparency and accuracy of a system is frequently observed also for the intuitive reason that more data and inferential rules imply “more complex functions to be approximated” (Arrieta, et al. 2020, 100), and therefore greater accuracy in different circumstances. On the other side, this kind of complexity will tend to make the model more opaque and adding an extra step in favour of transparency is likely to decrease its accuracy. However this is not an inevitable fallout: if data are well structured, it is likely that greater transparency will not correspond to less accurate AI performance.

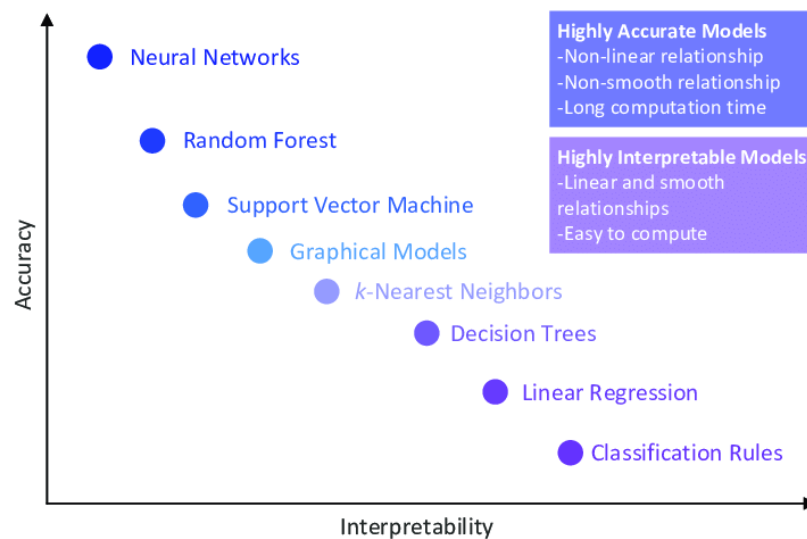


Figure 1. Accuracy and Interpretability trade-off (Morocho-Cayamcela, Haeyoung and Wansu 2019).

Finally, to think that understanding the functioning of an artificial system corresponds to visualising only its inner workings is deceptive. As Mike Ananny and Kate Crawford put it: “[...] rather than privileging a type of accountability that needs to look *inside* systems [...] we instead hold systems accountable by looking *across* them—seeing them as sociotechnical systems that do not *contain* complexity but *enact*

complexity by connecting to and intertwining with assemblages of humans and non-humans” (Ananny and Crawford 2018, 2).

The lack of transparency or trackability that AIs can suffer from can lead to accountability gaps. An alternative paradigm to transparency has recently been that of AI explainability (XAI), as a most effective and suitable form of keeping people informed about how the AI system works.

4.1. Explainable AI (XAI)

Apart from the inherent technical issues, however, there are doubts as to whether AI transparency alone is suitable or sufficient to support the detection and distribution of legal liability. Usually, for the purposes of establishing legal liability, in addition to mere exposition of internal mechanisms and data, it would be needed to have a more or less clear picture of what causes led to certain effects and to explain why. I have already pointed out in the previous section that for most complex AI models – e.g. artificial, deep, convolutional and recurrent neural networks as well as other ML algorithms – this task is not achievable by means of mere information disclosure (even where interpretable). The problem is that a trustworthiness gap of this kind or just undetectable biases risk cutting off very promising technologies that could take high-stakes decisions.

Where models are not as transparent as those of linear regression or classification, *post-hoc* rationalisation techniques – they can be grouped into the practices of the ‘explainable AI’ (XAI) – should be preferred. The main difference is that while for the former interpretability is a passive feature as they are transparent by design, for the latter additional explanatory techniques are required to make sense of how a prediction was obtained from an input. (Guidotti, et al. 2018; Gilpin, et al. 2018). There are two explanation targets: in the case of the so-called *global* explanation, the aim is to provide a comprehensive reconstruction of all possible decisions and their logic that an AI model makes; by contrast, in the case of *local* explanation the aim is merely to retrieve an explanation for a specific case predicted by the AI model (Ribeiro, Singh and Guestrin 2016).

Although some of them are model-agnostic (Ribeiro, Singh and Guestrin 2018), explanatory techniques are numerous and diversified according to the model they refer to; the most common ones employ

forms of reverse engineering and counterfactual explanations to reconstruct the rationale behind the prediction of a model where only input and output data are known, i.e. a black box models (Oh, Schiele and Fritz 2019). The core of these techniques is trying to explain the reason for an AI decision by demonstrating how the results produced by the model would have differed as the input changed. So, for instance, pursuant to a counterfactual explanation technique an AI system predicting the risk of criminal recidivism can be queried in this way: "if the defendant had previously been convicted of other types of offence, would the pre-trial detention measure have been imposed anyway?".

XAI has become such a central issue that the European Parliament has provided what appears to be a proper right to an explanation in Article 22 of the General Data Protection Regulation (GDPR). More precisely, the article states that where decisions based on automated procedures and profiling are permitted, the data controller: "[...] shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision". As you can imagine, the prescription is quite generic and its application is controversial among jurists and scholars (Wachter, Mittelstadt and Floridi 2017).

Ahead of the legal issues, what casts a shadow over explainability as a method of tracking AI decisions and human involvement are technical issues. Is it certain that *post-hoc* explanations are really able to represent the causal relationships of AIs' predictions? As some scholars argue, since such explanations are typically self-generated, they often prove to be inaccurate reproductions of the original computational model and strongly selective of the relevant information (Diakopoulos 2020, 205). Presenting some of the criticisms of the XAI, Cynthia Rudin observes that non-adherence to the original model is fairly predictable: "Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable)" (Rudin 2019, 207). As an example, in the case of Herbert, the trading agent for the purchase and sale of used cars, an unfair commercial practice was seen to result from the self-induced association between 'bait advertising' and optimisation

of the utility function. A *post hoc* explanation would be able to outline the inference rule thus obtained, but would not be faithful to the original model and the sort of human commitment in the misconduct (Rudin 2019). Hence, if explanation techniques do not contribute to making the original model interpretable, and in some cases even conceals it, then it is not an optimal strategy for distributing legal responsibilities of black box decision-making.

So on the one hand there is transparency, which finds objective obstacles in exposing all the data and information – and the price to pay when it does is high – and on the other hand there is explanation, which turns out to be a partial and somewhat unreliable reconstruction. From the technological point of view, both strategies should probably be pursued and the methods for making the systems interpretable and explainable refined. Yet, from the socio-juridical one, it is equally likely that a *soft* type of transparency already enables proper governance of AIs. After all, what use is an explanation if, as stated in section 3, liability can also be attributed regardless of causal links between damage and individual conduct? What seems important is rather a legally oriented standard of transparency, i.e. one that allows the legislator or policy maker to assess its adherence to the regulatory framework and the state of the art. In other words, whether the system is sufficiently understandable and when it is not, whether its use in a given environment is tolerable. Contextualising transparency helps to prioritise the purpose of disclosure, thus preventing indiscriminate exposure of information from lending itself to gaming, manipulation or privacy infringement.

Rejecting then the goal of full transparency⁴³ in favour of a *mild* one, aimed at rendering the computational model human-interpretable – at least at the starting blocks – to monitor AIs activities and to settle “easy cases”. Such as those cases where the malfunction is due to detectable (human) mistakes. If Herbert's software, for example, had explicitly provided for the possibility of bait advertising, or the acquisition of the malpractice was easily discoverable and preventable by the designer, then a mild transparent model would permit us to take the human agent as the accountable party. Anyway, for hard cases, where no predictions can be made about the future behaviour of an AI nor is it possible to

⁴³ Full transparency is also hampered by security issues such as gaming and manipulation practices or otherwise related to privacy as (Diakopoulos 2020, 213).

isolate the individual contribution in the socio-technical system, I still think the criterion of liability to be investigated has less to do with causal roles than with risks and costs distribution, likely inspired by economic criteria.

To conclude, the benefits of using methods of explanation or full transparency are still tentative, and for this reason it would be preferable to focus on a standard of mild transparency that nevertheless addresses AIs as socio-technical systems to exhibit the network of interactions between humans, computers and infrastructure.

In the current state of the art a certain degree of opacity is an unavoidable effect of complex AIs. Though, it is precisely these opaque systems that show greater effectiveness and accuracy in forecasting. It is probably not optimal to wait for technical standards of maximum transparency to be achieved, as this could take a long time compared to the short-term, if not immediate, benefits of using certain AIs. Accordingly, for the purpose of designing a liability regime (especially under civil law) a cost-benefit analysis should be carried out for each AI system in order to balance the social risk of licensing partially unpredictable artificial agents against the expected gains for the community. By the way, this seems to be the spirit behind the recent regulations on artificial intelligence at European level (e.g. Artificial Intelligence Act).

Integrating this type of assessment to design liability schemes is certainly not a new approach to legal science and is known to be particularly in use in the economic analysis of law. I will draw on some of the arguments of this approach notably in relation to the civil liability regime for AIs' failures; but first some brief considerations on criminal law liability for which the same approach may not prove as effective.

5. The status of AIs under criminal law

Although I shall privilege the study of the legal status of AIs from the civil law perspective, and the related liability regimes, some considerations are also necessary on the side of criminal law. The specific focus, as repeatedly stressed, is on the proposal to give the AI some form of legal personhood. I will not focus on a criminal law regulation of a specific legal system, rather I will make a general point by

touching only on the theoretical assumptions of criminal law shared by different legislations.

Despite civil and criminal law personalities being partially overlapping, the scientific discussion on the attribution of an autonomous personality to AIs is most advanced on the former one. The reason for this discrepancy is to be found in the peculiarities of criminal liability and punishment, which make the option of attributing direct responsibility to AIs highly controversial. Legal personality and liability are very much tied together in criminal law. Anyway, scepticism is not necessarily justified: after all, similar problems arose when discussions began on the criminal imputation of legal persons – e.g. companies (Coffee 1981). Nowadays there are no compelling objections to the recognition of the fact that also corporations are imputable for the commission of certain crimes.

For an entity to have a criminal law personality means to be the addressee of rules and duties governing its behaviour, the violation of which authorises the imposition of peculiar sanctions. Criminal law personality for AIs would presuppose that such systems possess the cognitive-behavioural attributes suitable for being subject to criminal provisions and that they will directly - if not exclusively - suffer the typical punishments. This is obviously just a minimal definition of what personality is about: throughout this chapter, I shall offer only a functional interpretation of legal personality (both civil and criminal), while I shall leave the more in-depth philosophical analysis to the second part of the thesis.

For the time being, and going in order, let us start with the following point: what happens if an AI engages in criminal conduct? A distinction must now be drawn between two situations: those in which an AI is used as a tool to commit a crime and those in which the production of the crime is only accidental. In short, the cases in which an AI entity is designed or employed for the specific purpose of perpetrating crimes and those in which it is not.

For both types of situations the default solution of the legal system would be to hold liable the involved human subject (whosoever). This seems obvious – even though but might not be – for cases where an artificial system has been deliberately programmed to commit a crime, but not for accidental cases. Anyhow, the assumption that the human subject is the only party capable of being blamed stems from two different, though ideally connected, types of issues: the first has to do

with the preconditions for criminal liability, such as the objective element (*actus reus*) and the subjective one (*mens rea*); the second has to do with the criminal punishment and its function.

As is well known, in order to be censurable under criminal liability, a person's conduct must be “reason-responsive” i.e. it must be guided by the agent's decision-making capacities according to different levels of awareness and volition (AV) (Edwards 1954; Fischer and Ravizza 2000; Duff 2007). The criminal behaviour must therefore be explicable or justifiable in the light of a corresponding mental state. Relevant mental states for criminal law range from premeditation and specific intention – higher levels of AV – to negligence, recklessness and malpractice – lower levels. When the subjective element is affected by pathologies or vices that impair AV, the conditions for the person's imputation are no longer met. But there are also residual and controversial cases in which the person bears criminal liability solely on the basis of the causal link (i.e. strict liability).⁴⁴

The *actus reus* does not pose any particular challenge to us: this requirement, which concerns material commission or omission conduct, can be satisfied even where the associated mental state is missing. As a result, there is no struggle to consider it to be present even for artificial entities. Conversely, the fulfilment of the *mens rea* requirement is more controversial and is often taken as the breaking point between human and non-human animals (and for personality attribution purposes). To what extent can an artificial intelligence system be deemed to comprehend and be willing to commit an offence? There is no doubt that an AI has the capacity to store information, perceive the external environment, and exploit this information to engage in a certain reasoning and course of action (Russell and Norvig 2016, 37). However, as mentioned above, the subjective element is also composed of the volitional element, albeit with different nuances. This drives us back to the point of intentional agency. I will deal specifically with this profile by focusing on the hypothesis of direct criminal responsibility – and therefore personality – of AIs. At present, just consider that the first relevant divergence arises around the *mens rea* element.

The second reason for mistrusting the idea of making AIs addresses of criminal norms concerns the modalities and purposes of the sanction.

⁴⁴ Some interesting discussions on the suitability of strict liability for criminal sanctions: (Gerber 1974; Simons 1997).

It could be assumed that the impossibility of implementing and enforcing the emblematic criminal sanctions against a “machine” – take the restriction of personal freedom as a strong case – also prevents some of the most important functions of punishment from being pursued/served. In particular, while the incapacitating and rehabilitating functions might hold – e.g. the offender system can be banned or reset and improved to avoid making the same mistakes again – the deterrent and retributive function (both specific and group) would risk fading away. Of course, this postulate is a corollary of the belief that because of the cognitive inability of AIs, who have no moral perception or cognition of their own interest, they would not even recognise the disincentives not to break legal norms. Scepticism about safeguarding the retributive function is not dissimilar, as the punishment of an artificial agent could be (socially) considered not proportionate enough to the offence (Hallevy 2014, 185). By the way, I doubt that these reasons relating to the function of the punishment and its modalities should worry us too much, but I will tackle this question again in relation to the hypothesis of direct personality and responsibility of the AIs.

As previously stated, we should distinguish all those circumstances in which an AI is programmed or used for criminal purposes from those in which the crime is a – more or less – unexpected and unpredictable consequence of AIs’ intentional agency. Within these different scenarios we can imagine at least two different set of regulatory instruments: one that contemplates the legal personality of AIs and one that does not.

5.1. AIs with “licence to kill”

Let me start with the scenario where an AIs is programmed, trained or employed for the purpose of committing crimes. Certainly this may be a ploy by which human agents try to shield their participation in the crime. If AI entities are denied any kind of legal personhood and treated as mere tools, then the only responsible parties are to be found within the ranks of the humans involved. In line with this scenario we find the proposal of Gabriel Hallevy: the “perpetration-by-another liability model” (Hallevy 2010, 10).

Hallevy's first model is built on the assumption that AI entities do not have ontological characteristics comparable to human beings; however, it neither takes the opposite extreme, equating them with any other technological artefact. He suggests that AIs should be regarded as

individuals with limited cognitive capacities - e.g. mentally incompetent or children - who are capable of performing material activities with a minimum degree of autonomy (at least the most advanced one). Yet, the criminal liability arising therefrom falls exclusively on the perpetrator who uses the incapacitated person as an intermediary. A split is thus created between the executor of the material conduct (*actus reus*) and the person in the corresponding mental state (*mens rea*).

In the hypothesis we are considering, the subject holding the relevant mental state may be the programmer who designed the algorithm in a way that it would commit a crime – e.g. a war crime for autonomous weapons – or who instructed it with intentionally distorted (maybe biased) training data; alternatively, the subject to be held liable may be the ultimate user who, even though the system was not 'licensed to kill', employs it with that purpose by giving instructions that the AI is prone to follow. Thus, these parties are those who plan the criminal conduct – i.e. perpetrators – while the AIs figure as innocent perpetrators.

But what if the AI entity has broad legal personality⁴⁵ – perhaps justified by the existence of cognitive competence – and is at the same time programmed to perpetrate crime? The two things are not contradictory: it is quite possible for a machine to be programmed to, say, commit financial fraud, but also engaging in autonomous reasoning to do so. This being the case, one could even think of a form of concurrence of legal subjects in the offence (when the offence is a single one). The artificial entity would be blamed for contributing to the commission of a criminal activity that was not specifically contemplated by the initial instructions of the designers or for doing nothing to prevent it. This borderline eventuality, though, leads us towards the second type of situation.

5.2. Unexpectedly criminal

The other type of situation is where the AI system is not deliberately programmed or used as an instrument to commit a crime. The “Perpetration-by-Another” model, as Hallevy himself acknowledges, would not accommodate those situations where an AIs "decides to commit an offense based on its own accumulated experience or

⁴⁵ As I will argue below, there are different forms of legal personality depending on the extent of the associated subjective legal positions (i.e. active or merely passive).

knowledge” (Hallevy 2010, 14). As a consequence, Hallevy holds that where there is neither intention nor knowledge to commit a crime on the part of the human subjects involved, but the AI entity has nevertheless caused one, another model should be adopted: the “Natural-Probable-Consequence Liability Model”. The AIs' misconduct in such cases may be due to accidents (f3): e.g. an untraceable bug in the data or in the code, or both, or self-learned patterns of conduct. *Mutatis mutandis* the case scenario of Herbert the electronic trading agent would probably fall into this set of situations, although in that case the offending conduct did not constitute a criminal offence, but only an unfair commercial practice; but we might also imagine Herbert taking up the purchase and resale of stolen cars and thus committing the crime of receiving stolen goods. Here again, I think it is worth imagining what would happen if the system were or were not accorded any form of legal personality.

Without any form of legal personhood to AIs, attention has only to be paid to the type of involvement of the human agent who, unlike in the previous hypotheses, has neither intended to commit a crime nor planned to use AIs for this purpose; in short, the *mens rea* component understood as intention is missing on the human side and it is assumed to be missing on the AIs side too. However, this cannot leave a gap in liability but rather forces a change in the type of subjective element by lowering the standard of AV. According to this model, the sufficient mental state to bring about a liability of the human subject involved is negligence: “Programmers or users are not required to know about any forthcoming commission of an offense as a result of their activity, but are required to know that such an offense is a natural, probable consequence of their actions” (Hallevy 2010, 17).

Briefly put, the developers and users of AI systems are liable if they have done nothing to prevent the reasonably foreseeable offence from happening. And that also applies, according to Hallevy, when there is some human criminal intent behind the use or design of an AIs, but then the system deviates from the initial instructions and commits a further offence: i.e. the human party will also be liable for negligence for the second offence. Although this obviously shifts the debate on what is reasonably predictable and what is not, the model seems to be suitable for a range of adverse events vitiated by negligence.

Cases where the programmer may be negligent are those where, for instance, the presence of incomplete, erroneous or biased data in the

dataset with which the algorithm is trained would be detectable according to the average operator (be it a programmer or a user); and the same applies for information which has been collected from untrusted websites. Alternatively, the human party may be negligent those times when an AI's decision required authorisation for being performed and the danger was detectable and avoidable (it is assumed that the system is not fully autonomous). In Herbert's case scenario, the programmer might be negligent if the goal of maximising profit was not limited in any way in order to avoid unlawful conduct at the design stage; or if the unfair commercial practice was not stopped immediately once it began to become identifiable, and so on.

5.3. Criminal personhood and direct liability

Against this background, what difference does it make if the AI has legal personality under criminal law? This presupposes that AIs are - at least potentially - able to fulfil both the objective and subjective elements of the offence and consequently to be directly liable for crimes. This is what Hallevy calls the "Direct liability model" (Hallevy 2010, 21). However, a further implication is that it would be possible to enforce the appropriate - i.e. proportional - sanction to the offence on AIs without undermining the very function of criminal punishment.

5.3.1. Mens rea and moral stance

The difficult part is to ascertain the occurrence of *mens rea* in an AI. What tends to matter for criminal liability, as we have seen, are the factors of awareness and volition. With respect to the former, we might first pose the epistemological problem of understanding what we are preaching when we say that someone *knows* something. According to the traditional conception an entity has knowledge if it has a justified and true belief (JTB).⁴⁶ It is also widely known that this definition of knowledge is subject to attacks and counterexamples of different kinds as to the sufficiency of the three JTB conditions to establish what is knowledge (i.e. the Gettier cases) (Gettier 1963). It is beyond the scope of this thesis to analyse this epistemological dispute, but it is still

⁴⁶ For an overview of the theory of knowledge and an introduction to the JTB view see, among others, (Armstrong 1973; Chisholm 1977; Nagel 2014).

important to find out which conception of knowledge we are happy with in relation to our question. Seeking to adopt an ecumenical approach, one will notice that the lowest common denominator of knowledge theories is the belief condition (Chopra and White 2011, 73). In short, that an agent perceives and comprehends the environment in which it operates (i.e. it has a *representational state*). Hence, the question to be faced is the following: what does it take for an AIs to have a belief? In some ways, answering this question brings us back to what was said in the previous chapter about agency and intentionality. In fact, here too the alternatives fall into two major families: the first is the realist one, according to which belief is that state of mind which corresponds to, for instance, the cognitive processes by which information is stored and instantiated (e.g. computational theory of mind); the second, on the other hand, is of the interpretivist type and thus focuses more on the attribution of knowledge according to the way in which a system or entity carries out certain actions. Within the latter we may locate Dennett's account based on the inference of the best explanation and prediction (Chopra and White 2011, 74). Already in the first chapter (Ch. 1, Sec. 5.3.2.) I tried to argue that the latter approach has significant advantages and, most of all, seems more functional in the legal context. Either way, both approaches - at least in the computational version of realism - allows us not to cling to a single model of "hardware" for *mens rea* evaluation, but rather to the "software" through which the information is processed. As Rohit Parikh theorises, in fact, social entities and social phenomena can be conceived as computational mechanisms – the so called "social software" – where information is stored, processed and implemented (Parikh 2002). In this sense, knowledge can also be attributed to a group of people. Collective knowledge, as also Game Theory aims to explain, will be used to coordinate and possibly cooperate to achieve certain goals (and equilibria). Generally, this kind of knowledge makes those in the group behave differently from how they would have done individually. This interpretation, over-simplified here, reinforces the analogy between AIs with corporations - which are liable in criminal law - and provides an extended reading of what knowledge is.

This being the case, when would we say that Herbert the trading agent is aware of something? "An agent's belief corpus is taken to be the set of propositions the agent is committed to [...]" (Chopra and White 2011, 77). Paraphrasing Samir Chopra and Laurence White, Herbert

knows a proposition p – e.g. a state of the world - iff: (1) p is true; (2) Herbert has access to the relevant information to know the content of p ; (3) Herbert can make use of the information of the content of p to perform his function and pursue his goals; (4) Herbert has access to all this information through reliable cognitive processes (i.e. non-accidentally).

From a technical point of view, it is perfectly plausible for an AIs to fulfil those conditions. An AIs can in fact perceive the external environment and produce descriptions of it via sensors and then filter out task-relevant data (Lagioia and Sartor 2020). The perception of the environment is not a unique skill of robots with sensors, even a trading bot like Herbert can do this thanks interaction with other market agents. Anyway, it is also technically plausible that AIs process through inferential processes the information gathered through context representation in order to select (generally) the course of action that best maximises the utility function. This indicates that artificial agents make functional use of the information they hold and plan their future behaviour on the basis of it (projection in fact can be seen as a key feature of awareness).

For the volitional component of the subjective element, as described in Ch. 1, Sec. 5.3, intentional states can be attributed to AIs as resulting from the enactment of representational and conative states. To sum up, an (artificial) agent has the *intention* to accomplish X when it associates a planning and deliberation process with the pursuit of a goal Y on the basis of its knowledge. The targets an AIs pursues can be of a different nature – long or short term, conservation or change of state (i.e. delta goals) – and are formalised in a utility function the system will try to maximise through the course of action that has more chances of doing so. The course of action with the highest probability of success is chosen from among other alternatives through inferential and predictive modelling.

On closer inspection, the mechanism just described is precisely the one relevant to ascertaining the intention for criminal liability in general: to determine whether a person had the intention, for example, to kill the partner, attention is paid to the fitness of the criminal plan, as well as its practical implementation, to achieve the criminal objective. Admittedly, there are several shades of intention that play a role in criminal blame: premeditation, malice (or specific intention), recklessness and negligence are the most important. What varies between these volitional attitudes is

the intensity of the awareness, the desire for the criminal consequence to occur and the inclination to foresee the outcome and side-effects. But they all can ground criminal liability, also when unintended effects are not part of someone's reasons for action (Duff 1990, 74). If, for instance, a neo-Nazi sets fire to a Jewish library as a political gesture, but ends up killing five employees, she will be held liable for these murders for being aware and accepting the risk of the side effects of her conduct (recklessness). The same holds true where the consequences of the conduct are neither predicted nor accepted, but should have been because they were reasonably foreseeable (negligence).

The hypothesis that an AI acts with some kind of criminal volition is supported by the evidence that they perform deliberative and planning skills. Volition can be directed towards criminally punishable conduct, chosen among alternatives of which the AI has knowledge. If a system meets these requirements, and therefore the subjective element (together with the objective one) of the offence, the possibility of attributing direct criminal liability to AI takes hold.

Yet, some might argue, this proves nothing: there are other entities that fulfil *actus reus* and *mens rea* in relation to certain behaviour and yet are not held liable under criminal law. The obstacles that arise for some of these entities, think of children or mentally ill, would also arise for AI, namely a lack of the moral sense: the capacity to establish what is right and what is wrong (Hallevy 2010). Such a critique would be superficial for a number of reasons, but let us assume that it is justified with regard to the cognitive competence being required by criminal liability. That being the case, artificial intelligence agents do not seem to be in principle excluded from normative reasoning or moral judgment, i.e. to be normative agents. If moral knowledge is a specific set of beliefs and intuitions about what ought to be and what ought not to be, we can certainly teach it to AI. It follows that AI algorithms can account for what is forbidden and what is permitted in their deliberative process. Just as with the intentional stance, the behaviour of normative agents can also be predicted and explained in the light of the moral stance (Chopra and White 2011, 178).

At this point, a further objection could be raised since AI having normative constraints does not imply AI deliberating normatively. This objection too, however, should be rejected, since some artificial agents can, according to the context and under particular circumstances, choose either to comply with norms or to violate them on account of more

important objectives (i.e. deliberative normative agents) (Castelfranchi, et al. 2000).

Thus for an artificial agent a normative belief - moral or legal - can be part of its reasons for actions and contribute to plan its conduct, despite not being then adhered to. A normative agent may for instance decide to violate privacy rules when it fears that the agent it interacts with - perhaps contractually - is actually malicious software (Castelfranchi, et al. 2000).

To conclude, the subjective element of the offence can be assumed to be satisfied by AIs, particularly those in which normative deliberations are implemented. Moreover, it should not be precluded that some of the justifications and exculpations – e.g. insanity, self-defence, duress etc. - may be accorded to artificial agents who break the law. Or, for instance, a self-driving vehicle hitting a pedestrian in order not to hit three – the ethical dilemma portrayed by the thought experiment of the trolley problem (Foot 1967) – might benefit from the necessity defence.⁴⁷

Bearing this in mind, cognitive attitudes may not be sufficient to allocate direct criminal responsibility, which also accounts for the socio-legal reasons embedded in punishment and its function.

5.3.2. Punishment and its general purposes

The second requirement to be met in order to make the option of direct criminal liability of AIs sensible is to provide for an adequate punitive system, i.e. capable of pursuing the typical functions of criminal sanctions. In other words, how is it possible to sanction an AIs and preserve the general purposes of punishment? As with the general concept of legal personality, I believe that a functionalistic interpretation of what constitutes (criminal) liability should prevail. Of course, this exercise is by no means new to legal science since similar problems have been raised in relation to the criminal liability of companies. In that case, through analogical application and adaptation of criminal sanctions specific to individuals, it was possible to conclude that corporations were punishable (Hallevy, 2014, 212).

Commonly recognised as the four main purposes of punishment are: (1) incapacitation, (2) specific and general deterrence, (3) retribution and

⁴⁷ A large number of works have also been devoted to the ethical dilemmas posed by artificial intelligence. See, among others, (Wu 2020; Davnall 2020).

(4) rehabilitation (Alschuler 2003; Frase 2005). These objectives are ensured by means of sanctions - principal or ancillary - that affect individuals and their assets in various ways: fines, community service, suspended sentences, public office ban, suspension from exercising a profession, imprisonment, the death penalty, etc. To test the hypothesis of the direct liability of AI entities it must then be investigated what the "electronic" counterparts of such sanctions might be. Among other things, it would be appropriate not to subsume the sanction entirely under the fine, unless one wishes to re-evaluate entirely the concept of criminal punishment itself; which is not, by the way, a foregone conclusion, as will be seen when discussing civil liability.

(1) The incapacitating function is intended to prevent the offender from continuing to cause harm to society. This objective is typically pursued with measures restricting personal and operational liberty: e.g. death penalty, imprisonment, public service, disqualification from working in certain contexts (e.g. public employment). It is sometimes forgotten to consider that precautionary measures too - e.g. arrest, house arrest, seizure, etc. - have an incapacitating effect, even if they are intended to prevent the trial and the proper exercise of judicial functions from being compromised. Anyway, if it is true that the purpose of these sanctions is to incapacitate the entity from perpetuating the harmful effects of its conduct on society, then it remains possible for such sanctions to be applied to AIs. The latter may in fact be deactivated altogether or banned from exercising their functions temporarily. In some cases, the deprivation of liberty might only concern limited contexts and interactions: for instance, Herbert might be suspended from selling, but not from buying cars; or he might be deactivated in his function as a negotiator, but kept operational as far as the search or filtering function is concerned. Under all these measures, incapacitation seems to be preserved.

(2) Deterrence is undoubtedly one of the most important purposes of criminal punishment and is aimed at creating disincentives both for the general public and for individuals to engage in criminal conduct. Any sanction (also of civil law), as long as it is proportionate and appropriate, should aim at disincentivising illegal behaviour and, at least for human beings, the highest disincentive value is assumed to be the death penalty. Equally, we should expect punitive modes against AI entities to achieve this effect. Since, as has been noted above, AIs have tasks, intentional agency, and respond positively to rewards and negatively to penalties

(e.g. reinforcement learning), the threat of limiting some of their drives through punishment can produce a deterrent effect. Deterrence can be achieved either through measures restricting (operational) freedom or through financial penalties: an AI algorithm will estimate as disadvantageous to repeat the behaviour that has caused a slowdown in the realisation of its utility function or an economic loss. Moreover, as Hallevy suggests, if an AI has no assets with which to pay the financial penalty, it can be converted into wasted working time (Hallevy 2014, 226). Whether such sanctions will also have a general deterrent effect cannot be ruled out: if AIs can share information and experience, maybe through cloud computing or IoT networks, then disincentives to certain behaviours may become a matter of shared knowledge. As for human agents, monitoring whether or not AIs abide by the norms of conduct when made aware of criminal sanctions will help to verify the efficiency of deterrence measures.

(3) Retribution is the oldest of the general functions of punishment. Traditionally, it should ensure that sanctions produce adequate suffering to mitigate feelings of private revenge. Retribution is often delivered by incapacitating sanctions and the same would hold true if they were employed for AIs; but it might be objected that, at the end of the day, those who really suffer from the incapacitation of AIs would be the human subjects involved. Yet, this argument seems a slippery one since there are not any such apprehensions with regard to criminal liability of legal persons: “When a corporation pays the fine, its sources are absent for its workers, directors, clients, etc. This absence is part of paying the fine, regardless the identity of the offender, therefore artificial intelligence systems are not unique in this context” (Hallevy 2014, 227).

(4) Finally, probably the noblest of the general aims of criminal punishment: rehabilitation. Similar to deterrence, the scope of rehabilitation is to induce the offender to better behaviour and, as a consequence, to prevent recidivism. However, the procedures expected to ensure this effect profoundly differs from deterrence: they mostly consists of social reintegration measures, community works and individually customised programmes. What is provided is thus a form of re-education of the offender so that she can compensate the society in some way. The targeted effect is the eradication of the underlying motive for the deviant behaviour through services of social utility. It is widely perceived that harsher measures such as imprisonment are often not at all conducive to reintegration. Indeed, it could be argued that prison,

instead of rehabilitating, produces new forms of delinquency (Foucault 1975). But this aspect is certainly beyond the analysis of this work. Anyhow, what sanctions correspond to community service or probation in view of the rehabilitation of a criminal AIs? As mentioned, the aim is basically to prevent criminal conduct for the future without radically incapacitating the offender. AIs could, accordingly, undergo structural revisions or corrections without being deactivated and thus continue with limited and supervised functions. An AIS might also serve a public purpose - e.g. community service – and compensate society with its own expertise (provided dysfunctions have been corrected) (Hallevy 2014, 223). Herbert, for instance, could be provisionally turned off for its private use and offer its services free of charge to anyone who wishes, or its algorithm could be temporarily diverted to other functions, e.g. a platform to manage the sale and purchase of public transport vehicles.

5.3.3. Direct criminal liability: uses and benefits

Against this background, the hypothesis of conferring legal personality under criminal law to AIs, and therefore holding them directly liable, seems abstractly consistent with some of the prerequisites of criminal punishment. Yet, it is not enough: it is certainly appropriate to question what practical - e.g. regulatory - motives require this hypothesis to be followed.

If an AIs is given legal personality, then following the commission of an offence it may happen that: the AIs can be jointly and severally liable with the relevant human agent (user, owner, developer, etc.) or it can be held liable alone. The first case may occur either in situations where the AI system has been explicitly designed for criminal purposes or in situations where there is negligence or malpractice – e.g. in programming, control, supervision – on the part of the human agent. In both circumstances, the AI would be held liable for not doing anything to prevent the wrongful conduct from occurring, or for failing to remedy the negligent human agent's faults.

When is it instead possible to consider AIs as the only responsible party? We can infer the actual advantages of granting legal personality status to artificial agents by answering this question. As I have already anticipated, in the course of this thesis I give a functionalist reading of legal personality, i.e. as a way of optimally distributing the imputation of legal positions (rights, duties, liberties, liabilities, competences etc.). It

follows that the added value of having a new subjective point of imputation, as in the case under study, consists in facilitating regulation and increasing the comparative utility in relation to the ordinary distribution of rights, duties, powers and responsibilities.

So, what legal imputation issues does the “artificial personality” solve? Is this a more efficient solution than one in which an AI is treated as a mere artefact? I believe that, as far as criminal law is concerned, autonomous personality can emerge when the offence is the result of sufficiently autonomous conduct on the part of the AI to the extent that it cannot be traced to the intention, the mistake or negligence of the human party; or of a single, identifiable human party (i.e. the many hands problem). AIs might be held as the only responsible parties for those behaviours, maybe resulting from the self-learning capacities or inferred by data, which cannot be reasonably foreseen or controlled. In short, that class of events which have previously been labelled as accidents (f3). To oversimplify, The rationale is that those facts are not the “natural-probable-consequence” of the human involvement. Some of the essential prerequisites for responsibility risk vanishing in this way, namely the epistemic condition – i.e. a sufficient degree of awareness of what is happening – and the control condition - i.e. control over the actions and deliberations undertaken.

Of course, it is a matter of technical concern how far a given event was foreseeable or controllable by humans (and by which specific human party). My impression is that the greater the autonomous agency of AIs, the greater the frequency of cases in which identifying human negligence is difficult or almost impossible. Indeed, higher deliberative and operational autonomy, as we have seen, often corresponds to a stronger inner opacity of the system (e.g. DNN).

I might be wrong, but the opposite scenario, i.e. holding the involved human parties responsible for all wrongdoing, does not seem to be the ideal solution both from an economic point of view - e.g. it might discourage their production and use – and from a moral point of view – e. g. fairness and proportionality. However, this does not necessarily exclude the possibility that some responsibility on the part of competent human subjects persists, but recognising two different centres of legal personality makes it possible to diversify the chain of legal duties and liabilities. It should not be overlooked that deeming AIs as subjects of criminal law may be useful for those situations where the programmer of a crime-causing system is not a human being but an AI entity itself

(Hallevy 2014, 32). And we know that the hypothesis of AI algorithms building other AI entities is not unrealistic, since automated machine learning (AutoML) projects already exist.⁴⁸

I shall conclude here my remarks on the personality of criminal law for AIs. This regulatory solution seems practicable in principle in the area of criminal law since the requirements of criminal liability and punishment could be met. In addition, this diversification of legal spheres would make it possible to limit the duties of human participants in situations where it would seem unfair or too costly to keep them liable.

I have not yet addressed the most relevant objections to this regulatory scenario, because I reserve the right to do so in conjunction with the objections to the same hypothesis applied to civil law. In fact, similar arguments will be proposed in support to the conferral of civil personality on AI entities; thus, arguments for and against can be partially combined.

⁴⁸ For an overview see (He, Zhao and Chu 2021).

III. The Status of AIs Under Civil Law

1. Civil law and AI

It is now time to address legal personality under civil law. As mentioned, being a person in the legal realm means being the holder of one or more legal positions, having both an active and a passive dimension (e.g. protection of integrity). Up to this point, and also in the following sections, particular attention will be paid to liability, which is just one of these positions and not all legal subjects properly hold it, as is the case with infants or mentally incompetent. Yet it remains one of the most decisive positions in determining whether or not to confer legal autonomy on a given entity.

In view of this, the prior of the argument I will defend in this thesis is that one of the main functions of legal personality, as shaped by the peculiarities of legal systems and practices, consists in the efficient allocation of legal positions, e.g. rights, duties, liberties, powers, immunities, and liabilities. And the choice of extending the area of legal personality to further entities can be driven by the expected utility resulting from a specific distribution of these positions.

A variety of criteria can be invoked to determine whether the allocation of legal positions to a subjective point of imputation is optimal or not: moral, systemic and economic are among the strongest reasons. The way of measuring the efficiency of the distribution of legal positions varies from case to case, and in some circumstances may be less relevant altogether. With regard to AIs, it has been seen that among the reasons justifying the attribution of autonomous personality the most

convincing relates to the (better) distribution of liabilities resulting from unexpectedly illegal and/or harmful conduct or side effect of intended behaviour. So, in this case, an optimal allocation of legal positions may result in an efficient management and distribution of the social risk and costs associated with the use of AIs, ensuring the incentives for production and further improvement of these technologies.

Liability contours, albeit crucial, are not the only ones to be examined: specific competences and powers may be implied by being a subject of (civil) law, such as, for instance, financial autonomy, the capacity to hold property, to make contracts, to sue and be sued. It will be seen that these prerogatives may also contribute to the pragmatic and economic rationale for attributing personality to AIs.

I plan to compare different ways of distributing legal positions, and their results, by discussing three conceivable legal statuses for AI systems: as cognitive tools, as parts of juridical persons, and as proper autonomous persons. I will present the pros and cons of these regulatory options trying to argue that the third one, which can occur in various combinations, may be a valid solution also in the context of civil law; primarily to cope with liability issues.

Before that, however, it should be pointed out that there are substantial divergences between criminal and civil liability that contribute to making the latter more suited to a functionalist and economic approach to personality.

1.2. Civil liability and its peculiarities

Civil liability originates in a *civil wrong*, triggering an obligation to compensate for the damage caused by the wrongful act.

The compensatory relationship may arise either between two subjects who have come into contact by chance or for those who were already in contact by virtue of a contractual agreement. Among civil wrongs, in fact, are both torts – injuries to person or property – and breaches of contract or of trust.⁴⁹ Civil wrongs are the sources of compensatory obligations in civil liability and consist, as in criminal liability, of objective and subjective elements. The objective elements include, of course, the committed act, the wrongful damage and the causal

⁴⁹ These elements are the lowest common denominator of civil liability in most legal systems, both common law and civil law.

relationship between the act and the damage; the subjective elements are, also here, intention and fault.

However, there are significant differences between civil and criminal liability. First of all, the general purpose: usually, civil liability is much more compensatory and risk-allocative rather than punitive. In other words, the focus is mainly on restoring the damage - whether pecuniary or non-pecuniary - suffered by the injured party through the compensatory obligation of the damaging party. This obligation often consists in the payment of a monetary sum, but may also correspond to an obligation *to do* something (the Italian civil code speaks of specific compensation in art. 2058). In any case, unlike criminal sanctions, which concern the conduct of the offender *per se*, what counts for civil liability is basically the damage.

This divergence has important repercussions on the subjective element: while criminal liability is mostly agent-focused, as shown by the fact that victims are often powerless in criminal proceedings, civil liability is focused on the victim's injured interest (Cane 2000). It follows that the role of intention for, say, tort liability is notably reduced. Negligence, on the other hand, remains central even though it can be reformulated in more relational terms: "Tort law explicitly seeks to balance agent-autonomy security, while not ignoring wider social interest. Thus the legal definition of negligence refers to the interests of the victim (the harm done and its probability), the interests of the agent (the cost of precautions), and the wider interests of society (the value of the agent's activity)" (Cane 2000, 553). This last point is clearly compatible with an economic approach to civil liability, i.e. where negligence is less linked to the psychological dimension to make room for an efficiency criterion (Coase 1960; Landes and Posner 1987). And so, negligence subsists when the cost that would have been incurred to prevent the damage was lower than the value of the damage itself given the probability of its occurrence.

The element of causation is also stronger for criminal liability than for civil liability. In fact, for the purpose of ascertaining the extent of the damage to be repaired, all the consequences stemming from the breach or injury will be taken into account without there being a strong causal link with the original event.

That civil law places less emphasis on some of these elements is confirmed by the existence of peculiar forms of liability, which are in some way fault-independent, like *secondary* and *strict liability*. The first

provides that an entity other than the person who materially committed the act must be held liable (thus breaking the typical causal chain). The second, on the other hand, provides that a person who is in a certain relationship with the thing that caused the damage is liable for the fact, regardless of any fault (intention or negligence). Sometimes a mixture of these two forms of liability occurs (e.g. product liability in Italian legal system).

It gets difficult to go into more detail now since there is no general theory of civil liability and each legal system may have its own peculiarities. However, some forms of liability are fairly common: e.g. there is a well-known case of secondary (tort) liability, namely vicarious liability. In this case, a person is liable for the acts of another by reason of a relationship of control, direction or supervision between two subjects, the superior and the subordinate (*respondeat superior*). The relationship may be either pre-existing or arising on the occasion of the harmful event. Cases of vicarious liability typically occur for the employment relationship, principal-agent relationships (agency law), parental liability and legal person's liability for torts committed by its members. According to the circumstances, the causal chain may be interrupted altogether or remain relevant because a wrongful act of the principal is detectable in the breach of the duty of supervision (i.e. *culpa in vigilando*).

On the other hand, recurring cases of strict liability are: dangerous activity liability, product liability, custodian liability, liability for harm caused by animals and few others. The common assumption to these forms of strict liability is that whoever performs certain activities, holds certain positions or owns certain goods/animals, accepts, more or less consciously, the risk of economic loss due to (almost) inevitable events. Discharging in these situations is much more demanding: e.g. in the Italian system the proof of prudence or diligence is not sufficient to be exempt from dangerous activities liability, but instead it must be proven that all appropriate measures were taken to avoid the damage (Article 2050 Civil Code); the exonerating proof is equally demanding in the case of the custodian liability, where the custodian will be released only in case of a fortuitous event.⁵⁰ Sometimes, forms of strict liability can also

⁵⁰ As Galgano points out, however, it is not obvious that the fortuitous case is less demanding proof than the one in force for liability for dangerous activities: if there is no fortuitous case or if the cause of the damage remains unknown, proof of having

be found in criminal law - e.g. liability for traffic offences - but these are quite residual cases.

What relevance do these profiles have for the regulation of artificial intelligence and its civil law status? While criminal liability required questions about the existence of cognitive/psychological attributes – both for the purposes of culpability and punishment – civil liability has undergone a process of objectification that is more concerned with how to allocate the social risks inherent in certain human activities. The role of fault in the attribution of damage is greatly reduced (sometimes eliminated) by shifting the allocation of risks and resources to an efficiency criterion assessed in the light of collective welfare or general utility. This assumption is particularly recurrent in the law and economics, where causation becomes secondary to the importance of allocating resources efficiently irrespective of who causes harm (and to whom).

Efficiency may be pursued according to different approaches: e.g. (a) it may consist in splitting up damages, as widely possible, in relation to both subjects and time; (b) it may adhere to the so-called “deep pocket” method, whereby damages must be borne by those entities most able to bear the economic burden; (c) finally, it may attribute damages exclusively to the activities that cause them (Calabresi 1997; Calabresi 1961).

Patterns of these kinds are widespread in civil law liability, as confirmed by the presence of insurance systems – social or personal, compulsory or optional – which distribute the costs of certain accidents among several categories of people. Insurance systems whose service may also be suited to managing liability costs for AIs (Pütz, et al. 2018).

It goes without saying that a rather functionalist interpretation of civil liability applies here. And one of the functions of civil liability - especially the tortious one - that has been consolidated over time is therefore that of reducing the cost of accidents. Which means, as Guido Calabresi has famously argued, reducing the number and severity of accidents (preventive function) and the cost they cause to society at large (reduction of secondary effects).⁵¹ Legal personality in the functionalist

adopted all the appropriate measures to avoid the damage will not be sufficient to exonerate (Galgano, *I Fatti Illeciti* 2008).

⁵¹ These are the three sub-categories of liability accident cost reduction functions according to Guido Calabresi. In particular, the Italian-American jurist argues that

reading I advocate, e.g. a mechanism for distributing and limiting personal obligations, is one of the devices that can implement this feature of civil liability.

If it were necessary to clarify it, this does not mean that the psychological states, like intent or negligence, never have a place in the tort law system. As we shall see, they can also serve a significant function in regulating liability for damage caused by AIs. Yet, it should be stressed that, by virtue of the objectification of civil liability, there are multiple solutions to be considered and some of them take a social welfare perspective rather than blaming or punishing only one of the parties involved. In the case of AI, this could prove very useful in distributing risks not only over individuals - e.g. only over designers or only over users - but transversally, pursuing an interest that is as shared as possible. From this perspective, the cognitive competence of the AI and the psychological contribution of human agents may play a marginal role in shaping the liability model.

Although there is no general theory of civil liability, or personality of civil law, it is still necessary to refer to some legal system. In this work, I will take the European system as a working hypothesis because it represents a very heterogeneous legislative context which, nevertheless, needs to find a regulatory synthesis on a fairly new issue. As we shall see, this partly complicates matters, but at the same time the scarcity of common legislation on the topic allows us to treat the phenomenon as something to be built almost from scratch.

2. The social cost of AI

To frame the accountability relationship, as well as broader AI governance, and to find out which regulatory strategies can enforce it in AI, I wish to adopt a sociotechnical perspective. This means not only considering the accountabilities of individuals – for conduct or decisions relevant to possible AI failures – nor simply the accountability of the AI itself – e.g., whether it has a sufficient level of transparency or trustworthiness – but the whole context of which the AI is a part; thus including rules, tasks and practices (Trist and Murray 1993). This approach has sometimes been conceived as a shift from the traditional

there is also a third function that is concerned with ascertaining that the cost reduction project is not economically inefficient. (Calabresi 1997).

human-in-the-loop paradigm to a new one, the society-in-the-loop. This paradigm is based on the idea that AI governance must be inspired by a new type of AI-mediated social contract, which requires the participation of different stakeholders in supervision procedures and in balancing the values, benefits and costs involved in the use of algorithmic systems (Rahwan 2018).

In any case, the complexity of sociotechnical systems often makes it impossible to assign a single cause or responsibility for each error or failure. Sometimes, tracking such events is completely uncertain. From the regulatory point of view, this incentivises the adoption of a risk-based approach, in which regulatory burdens, including those on accountability, are proportional to the value of the interests at stake and the social costs of AI employment.

Prior to understanding how to distribute liabilities, it is then necessary to define what the social cost of AIs consists of. From the technical point of view, as already seen, there are several sources of risk and drives that can lead an AI to produce unexpected and untraceable outcomes. Injured parties of these outcomes may be both owners (or users) and third parties. It follows that producers or designers of an AIs find it difficult to estimate the inherent risks of their activity and accordingly their general utility (net of costs); when the injured party is a third party, the same applies to users and owners. Autonomy and self-learning are among the main sources of this information gap. Uncertainty is exacerbated by the fact that potentially each AI system could behave differently on the basis of accumulated experience. And if this were not the case, because the models are all being updated at the same time, perhaps the damage would be even more irreparable. In short, these failures can be seen as negative externalities of AIs.

Still, if the error rate remains lower than that of human deliberations – although some forms of human-machine interaction are desirable – then adopting sophisticated though risky technologies has also benefits to be taken into account. Just to give an example, according to the American Cancer Society, AI has a 99% accuracy rate in reviewing mammograms and is 30 times faster at diagnosing than human doctors (Patel, et al. 2016). This has made it possible to address the problem of false positives by reducing unnecessary biopsies. In the short to medium term, while waiting for the human species to produce infallible robots - which would prompt other kinds of questions - it would probably be unwise to give up such powerful tools of prediction and diagnostics.

Rather, it seems desirable to determine the optimal level of risk given the benefits of AI technologies.

In order for the cost calculation to be accurate, a risk assessment seems also necessary, for which one can take into account the recent legal framework proposed by the European Commission which differentiates three levels of risk: unacceptable, high, (limited) and minimal.⁵²

With this in mind, it would be appropriate to treat (certain) AIs' failures as a matter of social costs and to identify how to distribute them efficiently. The task is then providing incentives to prevent accidents (*primary costs*) and to do so in the least costly way for the parties involved (*secondary costs*). An effective regulatory policy oriented in this way should be able to transform unknown risks into known (*ex ante*) costs.

For secondary costs, the allocation of damages may follow the criterion of fragmentation, which postulates that it is (economically) preferable to distribute costs among several parties or over time, rather than in one large sum falling on a single party. A variant of this criterion is the "deep pockets" approach, according to which the fragmentation of costs should not be generalised or indiscriminate, since it is more efficient to distribute them only among those categories whose economic and social condition allows them not to be substantially affected (Calabresi 1997, 60). Of course, intermediate systems can be found, so there would be for instance more or less generalised fragmentation but with a 'deep pocket' approach for fundraising.

Damage splitting techniques are, among others, insurance (social or private) and corporate liability. While the former involves a broad fragmentation among different categories of subjects, possibly through taxation, the latter distributes damages among those who can best bear the losses or those who can offload the cost of damages on the buyers of their products. Or it might be possible to make those most likely to cause accidents bear the heaviest financial burden, even in the context of social insurance. In the case of AIs, these parties would be the developers rather than the users.

⁵² Proposal for a Regulation of The European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts Com/2021/206, Final, Document 52021PC0206, link: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>.

Anyway, a liability regime that aims to efficiently distribute the social cost of certain activities must also pursue the goal of damage prevention. This can be done in different ways, e.g. by prohibiting a certain activity or making it more costly. Therefore, one of the concerns in designing the type of liability for unforeseeable damages caused by AIs is to create incentives for technology to be improved and accident frequency reduced. Again, the players who can take action to improve technological standards and thus be targeted for such incentives are mainly on the side of AI production and development. But it is not obvious that in designing liability rules the fairest and most efficient solution is to affect one category alone.

From a regulatory point of view, pursuing both objectives, i.e. precaution and efficient distribution, is not an easy task. The combination of strategies in which this happens might be seen as an optimal equilibrium within a kind of non-cooperative game (Brown 1973). It is generally assumed that law is a suitable instrument for such policy objectives. Law can strive for these objectives either through direct regulation, through forms of taxation (e.g. so-called Pigouvian taxes) or through liability law. Rules on civil liability can aim at equilibria of this kind in several ways: e.g. they may internalise negative externalities through strict liability systems, discourage harmful conduct through negligence-based liability systems, or they may treat liability rights as proprietary rights thus promoting their free trade (Cooter 1991). However, we still don't know if there is a "silver bullet" for AIs' failures. As is often the case, there is insufficient evidence to know where the optimum lies, and analogies with other artefacts are quite precarious given the technological novelty. Likely the best strategy is to address the problem in a gradual way, without anticipating technological development and observing the actual deployment of AI devices to estimate the real extent of the risk and consequent costs.

In this thesis, as opposed to providing regulatory solutions, I explore scenarios of possible legal status of AIs and argue that in the one characterised by the spread and sophistication of this technology, there are valid reasons to confer them some form of legal personality. This hypothesis, as we shall see, stimulates broader philosophical reflections.

3. AIs as cognitive tools

The first civil law status to be discussed is also the one that is closest to the current legislation, both national and European. In fact, it comes as no surprise that artificial intelligence systems are treated as tools, albeit very sophisticated ones, operated by human agents delegating cognitive tasks. Such a legal classification naturally has significant implications for liability. On the occasion of torts or breaches of contracts caused by AIs conduct, we will tend to be concerned with identifying the accountable human party. For this purpose we may adopt, as mentioned in the previous paragraph, different liability schemes. In the European context, which is the preferred legislative case study of this work (along with the Italian context at times), there is no common discipline governing liability for damages caused by AIs. Indeed, there is a lack of harmonised liability law in general among Member States, except with respect to product liability (Directive 85/375/EC), liability for breach of competition law (Directive 2014/104/EU) and liability for breach of data protection law (governed by the GDPR). As the assumption of this section is that AIs are cognitive artefacts, yet artefacts nonetheless, the European product liability framework seems the most suitable legal container.

This lack of coordination is accompanied by a domestic regulatory vacuum in the Member States, which do not directly address liability issues of AIs - as well as other legal profiles - preferring for the time being to use existing disciplines and liability regimes by analogy (where possible).⁵³

However, the main categories of civil liability cross over to the individual national legal systems and it is thus tempting to look at the regulatory alternatives that a hypothetical European legislator of the future might tap into when designing a common framework for AIs liability rules. For the sake of clarity, contractual and tort law should be kept separate.

3.1. Contract law

⁵³ Report from the Expert Group on Liability and New Technologies appointed by the European Commission: 'Liability for Artificial Intelligence and other emerging digital technologies' (2019): <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608>.

By now we know that artificial systems can take part in legally relevant interactions like contracts. A large part of the transactions that take place on the Internet, in online markets or on ordinary shopping websites, are conducted and finalised by somewhat autonomous software agents. Also in the case scenario illustrated at the beginning of the chapter, Herbert was the name of an AI agent in charge of the acquisition and sale of used cars and that, in order to be able to carry out its activity, must take part in lawful contracts.

AI agents can be delegated to sign contracts on behalf of and in the interests of their users, but the content of these agreements does not automatically adhere to their directives in a rigid manner. This is the inherent vice - and virtue - of AI, i.e. often the direct supervision of the user as well as her ability to predict the actions of the artificial agent may be lacking. In other words: “[...] neither the user nor the programmer are in such a condition to fully anticipate the contractual behaviour of the SA (software agent A/N) in all possible circumstances, and therefore to “want” the contracts which the SA will conclude. Even when the user is in the condition of making such a forecast, he cannot be required to do so, since [...] this would contradict the very reason for using an SA” (Sartor, *Cognitive automata and the law: Electronic contracting and the intentionality of software agents* 2009 , 278).

Thus, the problem repeatedly presented above arises again: are such contracts effective, whose intention is the contract an expression of, and who will be liable for any breach of it?

We are at present ruling out the possibility that AIs are genuine persons in law and thus the chance of acknowledging them as proper contracting parties. Other legal solutions have to be envisaged. For all those circumstances in which contract agents merely carry over the initial instructions, without adding anything to the contractual terms provided upstream by the human operator, there seems to be no major concern about contract formation and the status of such AIs: they can be treated as mere passive communicative tools (or conduits). No variance between the contractual intention of the human user and that displayed by the AI agent; and the counterparty may expect the content of the contract to reflect the actual user’s understanding. The doctrine of the unilateral offer might be applied here, as is the case with vending machines (Chopra and White 2011, 37).

However, often artificial agents deployed to make contracts cannot be equated with mere conduits because it may be the case that there is

no correspondence between the contract concluded by the AIs and the relevant mental state of the human one (Sartor 2009, 279). To some extent, AIs by definition hardly fall into this category. If, on the contrary, we were to treat them generally as communicative tools, we would end up committing users to contracts that were not intended with the content (re)shaped by the AIs, or that were not intended at all. At least this holds true for those AI systems that are sophisticated enough to modify their behaviour according to ML algorithms. An example of this is, again, Herbert's, i.e. although it has rules to abide by, the weight it assigns to the parameters used for transactions, as well as the commercial strategies it implements, can be inferred in a probabilistic fashion from collected data and may change over time.

So, we should reject the mere-conduit-view as a general guide: we are considering AI tools that can potentially integrate the epistemic and practical authority of the human party.

For all these situations, the problem of contract formation and related liabilities becomes more challenging. Is the contract valid even if it does not reflect the intention of the human user? There are mainly two alternatives. One can be based on the assumption that the general intention to use an AI agent for contracts, and the acceptance of the risk inherent therein, is sufficient in itself to validate the contract and to justify ascribing any legal effect to the human user regardless of original intention. In a nutshell, the fact that the user has consciously delegated contracting activities to an AIs implies she is willing to accept all the consequences, even those that are undesirable and not specifically ordered. This can become a scheme of strict liability, which can also allow the liable party to be selected on the basis of economic criteria (e.g. deep pocket), and which seems sustainable as long as AIs failures are few and insignificant. This approach, within certain limits, is actually sensible: if every operation had to pass through the user's consent and approval, some of the very reasons for using AI tools would disappear, as it is the case of high-frequency transactions.

Alternatively, in case of unintended contracts, we might resort to solutions that better protect the human user by enhancing the latter's original intent and censoring certain unforeseen deviations. To make this purpose workable it will be necessary to investigate who's the negligent party: e.g. whether the deviant behaviour of the AIs was foreseeable on the part of the human user; whether the user was able to prevent it with relative ease, or whether the only ones able to do so were the developers

or producers. However, these notions, which seem to recall a kind of vicarious liability, do not often operate in contract law, where liability is rarely fault-based.

Yet, we are faced with very peculiar contract tools calling for the use of creative regulatory solutions. Some see the discipline of agency law as a possible alternative for this purpose (J. Fischer 1997; Smed 1998). Indeed, we should not be concerned with tracing the link between the intention of the human party and how it turns out as a result of being "manipulated" by the AI contract agent (Chopra and White 2011, 39). On the contrary, we could start to look at them as two potentially autonomous entities brought together on the occasion of the contractual affair in the form of a master-agent relationship. The ability of AIs to take part in a contract or change its terms without necessarily going through the acceptance of the human party, together with the possibility of interpreting the behaviour of such systems from the intentional stance (Sec. 5.3.2.), encourage such an approach.

These sorts of relationships are governed by the law of agency. Agency is traditionally a common law notion, but one that finds counterparts in various legal systems, and can be defined as the fiduciary relationship which results from the manifestation of consent between two parties, whereby the delegate (agent) may act on behalf of and under the control of the delegator (principal). Agency is then based on trust and tends to stipulate that the principal is always bound by the contract signed by the agent, even where there is no explicit mandate as in the case of implicit and apparent authority (Rasmusen 2004). Thus, an AIs may be delegated to act in the contractual interest of its principal and serve as an agent even in the absence of a contract of mandate signed by both parties. Besides, there could be no such agency contract, which normally occurs between legal entities, because it would necessitate the legal personality of the agent, which we are currently ruling out for AIs.

But we know that agency can work even if the intermediary agent has no legal personality at all. As several authors have observed (Kerr 1999), the scenario of contractual intermediaries without real legal personhood recalls the Roman law instrument of *peculium*: the dominus (akin to the principal) placed a sum of money at the disposal of certain agents lacking legal standing - mainly the slave or an unemancipated relative - so that they could perform legally relevant acts whose effects fell within the legal sphere of the former. Slavish application of the Romanesque discipline would lead to several problems, first of all for third parties who would

enter into contracts with electronic agents with the risk that the master would later reject them freely if she did not recognise their legitimacy (Dahiyat 2021, 64).

For this reason it is perhaps advisable not to deduce uncritically the legal status of a Roman slave, but rather to think of AIs as cognitive tools which, while not possessing any kind of legal personality or capacity to act, have the appropriate mental states to enter into a contract with a potentially divergent intention from that of the principal. And they would still be apt to compel the principal when the third party is induced to believe that the conduct is not vitiated by a mistake or an accident. In the absence of failures which are reasonably discoverable, the contract will not be voidable even if it goes beyond the authority conferred, because it would still be able to generate the expectation of *implicit* or *apparent* authority in the third party (Rasmusen 2004). In addition, the principal might not repudiate a contract even where she could have intervened to avoid the appearance of an authority, which was in fact not legitimate, but did not do so (i.e. *estoppel*). This discipline implies that subsequent breaches of contract and damages may also be ascribed to the principal.

If this is not the case, i.e. if the defect is recognisable by the third party and do not fall under implicit or apparent authority, the contract might be voided; in the event of contractual infringements or damages it could be then ascertained whether the failure was due to a programming error, a user-induced mistake or to the autonomous agency of the system (Sartor, Cognitive automata and the law: Electronic contracting and the intentionality of software agents 2009 , 279). In the latter case, the law of agency, given the absence of legal personality of the AI agent, would still keep the principal liable. But the same could apply to coding mistakes easily identifiable by the user (a situation that seems very rare).

To conclude, agency law seems a suitable solution for the contractual interaction between principals, AI contract agents and third parties. This is so for several reasons: it recognises the potential independency of the AI agent from the human user even within the framework of a non-person legal status (i.e. cognitive tools); it protects the principal from those blatantly wrong agreements resulting from the autonomous agency of the AIs or from human-driven faults; and it distributes the risk mostly on the subject that bears the lowest cost for the prevention of

damages (it may change depending on the problem but it is typically the principal).

On the other hand, the law of agency seems not to fit entirely with the characteristics of AI contract agents. For one thing, lacking an actual mandate and escaping the artificial agent's behaviour from the user's control, the relationship seems to be much more uncertain and hazardous than a traditional principal-agent relationship. Moreover, the analogy with agency would not cover those failures that are unpredictable for the principal and at the same time not recognizable as such by the contractual counterparts. In other words, it would leave out all the hard cases, i.e., situations of apparent authority, in which the consequences would keep on falling on the human party alone.

But even in the hypothesis of saving the principals from a greater number of failures, the law of agency would still create liability gaps. In fact, given that in the legal status scenario here in question no form of legal personality have been conferred to contractual AI agents, they would still lack a separate patrimony with which to insure their counterparts for any damages. As a result, there is either a return to a form of strict liability for the principal, or leaving the contracting counterparties more exposed to the risks of malfunction.

3.2. Tort law

Tort law is the area of law that governs civil wrongs, other than breaches of contract, by regulating the requirements of liability and how damages are to be compensated to injured parties. As already mentioned, tortious liability may arise from intentional or negligent conduct or from the mere existence of a legal relationship (strict tort liability). In short, liability for torts can be both fault-based and non-fault-based (Street 1999).

Questioning the tortious liability regime in the case of damage produced by AI systems requires tackling two types of issues, sometimes interrelated: the identification of the custodian and the causal link between the custodian's involvement and the harmful outcome. Typically, it is the victim who bears the burden of proof that the defendant's conduct was the reason for the damage. The likelihood of the victim being compensated for the damage depends precisely on the conclusiveness of the evidence produced.

Unfortunately, identifying the custodian and then proving its causal role in the damage caused by an AIs can be a tricky venture for the victim (and for the legislator too). Motives originate from the technical peculiarities of these technologies, namely the fact that the behaviour of an AIs relies on the combination of hardware (if any), the algorithm, the training data as well as data collected through self-learning capabilities. The subjects working on these AIs' components are generally multiple and diverse. To these subjects should then be added the individuals who physically operate and control AI tools, i.e. the operators or owners (here sometimes referred to generically as users). We know, moreover, that some useful information for improving performances can be shared or received by other devices - not necessarily AIs - thanks to cloud computing and IoT technologies. Updates to AIs can then be performed by different parties from those who developed or produced them.

Even if liability gaps do not necessarily arise - i.e. theoretically for any accident the liability of a human being can be invoked - uncertainty with respect to addressing the "better" party to be held responsible and proving causation may undermine the victim's access to justice and right to fair compensation. It is a choice of legal policy that must be evaluated in light of the costs and benefits to those who produce, those who use the technology and third parties in general.

This is particularly evident for fault-based criteria of tort liability, i.e. addressing the faults of the parties engaged in controlling the artefact (e.g. vicarious liability). Negligence of this kind can potentially exist at every level in the production chain of an AIs: it may affect codification, data processing, assembly, and, if these parties have done everything possible in accordance with the state of the art, then negligence may be transferred to the certification and marketing body. But negligence can also relate only to the way in which the device has been employed by final users.

However, because of the problems of reconstructing causation – especially for what have been called 'accidents' – it may be impossible to trace the single negligence or omission of control that was decisive in producing the damage. The same standard of care may be difficult to establish in the absence of a precise technological benchmark. All these aspects complicate the establishment of the defendant's fault both in terms of the burden of proof for the victim and in terms of the court's assessment of the violation. Ultimately, those forms of indirect liability, which hold the principal liable for the wrongdoing of another - in this

case AI agents - do not easily apply where there is no benchmark against which the misconduct of the human agent can be asserted.

As long as we intend to continue considering AIs as tools, these complications seem to suggest that non-fault-based liability systems should be preferred. At least for the set of uncertain damages, i.e. it is taken for granted that if negligence – e.g. breach of monitoring duty - is easily detectable there is no reason not to invoke the fault of the accountable party.

The alternative are then strict liability patterns, frequently used to place duties and liabilities for risky activities. Strict liability is sometimes further subdivided into *relative* and *absolute* liability: while in the former the perpetrator can be exonerated by justifying the damage by events totally beyond her control, in the latter the perpetrator will be liable also for unforeseen and unforeseeable damage.

Anyway, the main problem with strict liability is to determine which party should be held strictly liable in the first place. There are mainly two options, mirroring two different logics for distributing the social costs of AIs and with two dissimilar assumptions about which entity is best positioned to prevent, identify, and manage risk. They are both regimes of relative strict liability and therefore admit a number of justifications.

The first is to hold the human user (or keeper) strictly liable. Some consider the term *operator* more appropriate, because it would indicate the person who is in charge of manoeuvring the AIs and who derives the greatest benefits from their use. Yet, it must be kept in mind that being an operator is a variable role: “[...] ranging from merely activating the technology, thus exposing third parties to its potential risks, to determining the output or result (such as entering the destination of a vehicle or defining the next tasks of a robot), and may include further steps in between, which affect the details of the operation from start to stop”.⁵⁴ And that, in certain circumstances, the operator does not correspond to the ultimate user but to the: “central backend provider who, on a continuous basis, defines the features of the technology and provides essential backend support services. This backend operator may have a high degree of control over the operational risks others are exposed to.”⁵⁵

⁵⁴ Report on ‘Liability for Artificial Intelligence and other emerging digital technologies’, European Commission, 2019, p. 41.

⁵⁵ *Idem*, p. 41.

In terms of economic analysis of law, this liability regime would prioritise the so-called primary costs. Similar rules are present in existing disciplines, shared by several Member States, for example on liability for dangerous activities and are applied in the field of aviation, pharmaceuticals production or for animal liability.

The problem with strict liability designed in this way is that, as the degree of autonomy of systems increases, the operator's power of control decreases. In addition, the operator would have no particular right of access to the models and data on which the algorithm works – as the manufacturer has no legal obligation to disclose information to them – maintaining a position of unawareness of what is and will be going on.

3.2.1. Product Liability

Another strict liability model can be opted for which holds the producer always primarily liable, and therefore takes the view that she is best placed to manage the ongoing cost of employing AIs. In such a case, recourse can be had to product liability law which, as mentioned above, is one of the few liability regimes being harmonised at European level by means of a product liability directive (PDL). The directive makes the producer strictly liable for the defects of the products he has placed on the market, thus enhancing consumer protection. The rationale here is to give priority to so-called secondary costs, as the PDL considers that non-fault-based liability of the producer is the “sole means of adequately solving the problem peculiar to our age of increasing technicality, of a fair apportionment of the risks inherent in modern technological production”.⁵⁶

According to the current version of the regulation contained in the European Directive, products are basically all “movable” things and can be said to be defective if they do not meet the average safety expectations of consumers. Even if the injured party does not have to prove the producer's fault, the burden of proof of the damage, the defect and the causal link remains on the injured party. The producer may then invoke a series of exonerating circumstances (art.7, PLD) some of the most significant of which are: (a) the producer did not put the product into circulation, (b) the product did not have the defect at the time it was put

⁵⁶ Directive on Product Liability (85/374/EEC).

on the market, and (e) the state of the art technology at the time the product was put on the market did not allow the defect to be discovered (i.e. risk development defence).

There are undoubtedly some issues with these rules, which must be adjusted in order to bring the discipline of PLD more in line with the AI devices. First of all, it should be understood whether the concept of ‘product’ is able to contain AIs. Likely, the definition should be extended to include not only hardware but also software, and also allowing an AIs to be described in terms of the *service* provided, rather than just as a material artefact. A related issue is that the PLD framework maintains a technology-neutral approach. This regulatory strategy, as will also be later emphasized (Sect. 11), may not be functional with respect to AIs.

Next, as the advocated by the European Commission, the notion of defect and the burden of proof on the victim should probably be reformulated. In view of the complexity of technology and considered the inner opacity (real or engineered), the burden of proof for the plaintiff might be alleviated or maybe reversed, especially as regards the causal link as well as to the standard of technical adequacy required. The same conception of defects might be broadened to allow for compensation of those damages caused by defects that appeared *after* the product was put on the market. The reason for this last amendment is that, although at the time the product was certified or put into circulation there were no detectable defects, manufacturers continue to be in charge with updating the AIs and providing new data.⁵⁷

Finally, the recourse to the risk development defence should be restricted or denied altogether. The probability that producers will use this disclaimer to systematically evade liability to the detriment of victims is very high. In fact, as has been repeatedly remarked, unpredictability seems to be one of the constitutive traits of AIs technology. Nevertheless, it would remain available to the producer, as exonerating evidence, the proof that (a) the product has not been put into circulation, (c) the defect is due to having adhered to compulsory rules issued by public authorities and (d) the product was not manufactured by him for sale or other economic purposes (art.7, PLD).⁵⁸

⁵⁷ More precisely, these are some of the proposals for improving the product liability directive made by the group of experts appointed by the European Commission in 2019.

⁵⁸ Point (f) should also be added: “in the case of a manufacturer of a component, that the defect is attributable to the design of the product in which the component has

Such a revised regulation, harmonized among Member States, could ensure adequate protection of the victim's right to compensation and maintain a slight possibility for producers to exonerate themselves.

3.2.2. Absolute strict liability

Ultimately, absolute strict liability schemes may be tested too. They are generally implemented through compulsory insurance schemes or compensation funds. An example of this can be found in Italian tort law at the "Guarantee Fund for Road Victims" which provides automatic compensation for victims of unidentifiable vehicles. Such strict liability regimes are quite rigid, but this does not exclude that they may be practicable for damages committed by AI agents, perhaps on a subsidiary basis in combination with other liability regimes.

One way forward might be to adopt compulsory insurance with limits both on the amount of compensation to be paid and on the types of damage, e.g. not sufficiently covered by traditional forms of liability. One could, for example, use the capital capacity of the insurance for the so-called tragic resolutions that an AI takes (moral dilemmas are in fact the subject of debate for the ethics of AI); or in cases where the injured party does not have the appropriate amount of money to satisfy the claim for compensation. As for the compensation funds, these could be used to compensate for those damages produced by uninsured or unidentifiable AIs.⁵⁹

To conclude, while it is quite evident that using a single civil liability regime does not match all the challenges posed by torts of AIs, the compound of strict liability (relative and absolute) and fault-based schemes might be a sound policy. Generally speaking, the guiding principle, as seems to have been endorsed by the report for the European Commission, is to firstly hold responsible the best equipped party to recognize, control and assess the risk – also declaring it as a fault-based liability – and where this is not viable – due to reasons of non-traceability, danger of judicial dispersion and or insolvency – to resort to the typical instruments of strict liability.

been fitted or to the instructions given by the manufacturer of the product" (art.7 PLD).

⁵⁹ Report on 'Liability for Artificial Intelligence and other emerging digital technologies', European Commission, 2019, p.62.

4. AIs as (part of) juridical persons: company law proposals

A further legal status scenario, halfway between continuing to treat AIs as cognitive tools and assigning them some form of individual personality, is to integrate them into a distinct legal entity, e.g. companies or corporations. In this way, from the theoretical point of view, AIs would not properly accorded the status of independent legal actors, nor would they be merely objects under the direct control of the user, but would be conceived more as part of a shared cognitive system in which human and computer minds interact as part of a collective body. This would allow some of the features of legal personality to be derived, without radical dogmatic breaks: “It is worth noting here that this solution is more realistic since it may be easier to accept that a company has personality, intention, and other subjective states clearly more apparent than that of an electronic agent alone” (Dahiyat 2021, 76).

From the pragmatic viewpoint, the core of this insight is essentially to apply bundles of company law rules to AIs, e.g. the regulatory framework for limited liability companies (LLCs), as a way of settling their reception by the legal system and governing their powers and liabilities.

Replicating the legal status of the type of corporate legal entities would make it possible to extract the functional dimension of legal personality so as to ensure the ownership of legal positions directly in the hands of the AIs. Shawn Bayern, who is one of the proponents of such policies in the American legal context, stresses that: “The practical importance of this technique is that it allows software systems to achieve a very close surrogate for legal personhood”. (S. J. Bayern 2019, 25). Indeed, such a legal status allows AIs to perform the typical powers (and liabilities) of certain corporate legal entities: “such as entering a contract, owning property, suing, being sued, acting as a principal or an agent, entering into a general partnership, serving as a corporate shareholder, and so on” (S. J. Bayern 2019, 26).

This proposal is attractive primarily because of its “regulatory parsimony”, i.e. it would be possible to make use of existing company law disciplines without having to create a new legal status of the type of

'electronic personhood'.⁶⁰ Such regulatory solutions are clearly not the exclusive preserve of American law, but can also be found in various legal systems of the Member States and possibly be the subject of appropriate reflection by the European legislator (Bayern, et al. 2017). I shall here review some case studies that have already been analysed – such as the American and German LLC regulations – and I shall also investigate whether similar models are currently available in Italian corporate law and, eventually, if an harmonized discipline could be in the European law.

The gist of the argument is to build a juridical person around or housing an artificial intelligence system. This is not exactly the same as giving the AI its own legal personality that it can freely dispose of. Hosting an AIs in a juridical person might be justified by the will of one (or more) subject(s) and stakeholder(s) to allocate funds to carrying out an activity mainly managed by the AIs itself. As might be the case, for instance, for the creation of a non-profit foundation where it exists a specific interest in having the foundation's mission carried out solely by the computer tool “either to achieve redundancy in data replication and preservation or because at some point software may seem a more reliable tool than a traditional nonprofit foundation, this founder desires to establish a perpetual, autonomous foundation that captures and preserves information. The founder does not want to employ, and perhaps does not trust, individual people to manage the perpetual mission of the organization; instead, the founder wishes to commit certain resources to the organization’s software initially and then permit the software to act in an economically, functionally, and perhaps legally autonomous manner” (Bayern, et al. 2017, 137). One of the most common forms of corporate bodies are limited liability companies, which generally come into existence by virtue of an operating agreement regulating their functioning and organisation and which can easily be transposed into the code of an AIs. At that point, one or more members of the society would be able to confer to the AIs a detached property and assets with which the AIs could carry out its activities (commercial or otherwise) on its own. The human members would form the legal entity and then withdraw at a later stage. The legal status that an AIs

⁶⁰ This expression was used for the first time in an official document in the European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

would then acquire would not be comparable to that of a mere (cognitive) tool, but rather of a real LLC and therefore of an autonomous legal entity.

Furthermore, many legal systems also provide that a limited liability company may be a single-member company (also called one man's company), i.e. constituted by a unilateral act of a single founder, composed of one member only, and – like a normal LLC – still having its own assets.⁶¹ The sole founder of the single-member company may be either a natural person or a juridical one, even though usually the actual member must be a natural person. The relevant European framework is contained in Directive 2009/102/EC – replacing Directive 89/667/EEC – which covers both private and public single-member LLCs. Over time, there has been an increase in the use of sole proprietorships,⁶² mainly for smaller businesses, and this company dress could also be adapted to AIs used by individuals rather than in a shared manner (for which traditional LLC forms could be used). The existence then of an European framework and of analogous rules at the level of the Member States would likely facilitate harmonisation.

But there is a more extreme, albeit residual, form of company that is apparently suitable for hosting an AIs and which operates in the absolute absence of members. These are the memberless companies, once again with limited liability, contemplated by the German legal system. There is debate in German commercial law doctrine as to the validity of these forms of company, mainly grounded in the scepticism about the merits of a company without the organisational-decision-making structure - i.e. a multi-personal body – that tend to characterises collective entities. This concern, though, does not seem to be justified in the assumption that this corporate shell is tailored to AIs, which would instead retain deliberative capacity independently of the founders of the non-member company (Bayern, et al. 2017, 145).

Unsurprisingly, both traditional, single-member and memberless LLCs have advantages with regard to tort liability, as they distribute or limit the risk through the separation of the assets of the legal entity - in

⁶¹ Joint-stock companies may also be single-member companies, as provided for, for example, in the Italian legal system following Legislative Decree No 6 of 17 January 2003.

⁶² In the period from 2004 to 2013, there was an estimated four-fold increase in the number of single-member companies in the Italian corporate-business environment, according to data from the Union of Italian Chambers of Commerce ('Unioncamere').

this case the platform that would contain the AIs - and those of the members and founders. The European framework on limited liability companies is fairly homogeneous - see Directive (EU) 2017/1132 - and could allow for such a common policy.

Yet, there is another class of tools to be accounted, which concerns the earmarked assets. In the Italian legal system some of them have been introduced through the same reform that brought in the single-member company, and are known as “patrimoni destinati ad uno specifico affare” (Article 2447 bis of the Civil Code), i.e. assets earmarked for a specific purpose or business. Comparable schemes exist in other legal systems, sometimes they are called ‘Restricted assets’, or simply ‘Earmarking’, while other times they coincide with ‘Private foundations’ (see American and Swiss legal systems). Either way, what happens is that part of the assets of a juridical person are diverted to a specific businesses in a pre-defined and bounded manner, but without having to create a new subsidiary company. Similarly to the above-mentioned legal shells, also these earmarked assets may limit the liability of the members to the amount allocated in the assets. Under the Italian company law, the latter can never be higher than ten per cent of the total assets of the company from which it is detached and, unlike LLCs, it is not equipped with autonomous legal personhood, therefore remaining dependent on the juridical person that set it up. Compared to the LLC scenario, the advantage would mainly consist in the fact that the same company could allocate part of its assets to several AI-driven activities, without being required to create a separate company each time.

Hence, generally speaking, an AIs could be endowed with part of the assets of a juridical person, being bound to carry out certain activities falling within its social object, holding some degree of operational autonomy while being supervised by the board of members of the juridical person itself. But since such estates may in some jurisdictions have their own legal personality (e.g. in the Swiss legal system), the hosted AIs that fulfil their purpose will certainly be more autonomous and, as a consequence, acquire a status other than that of a mere instrument.

Not dissimilar conclusions can probably be drawn from the application of rules concerning the law of trusts, where the AIs may appear as the trustee to whom property or sums of money are given.

All the company law solutions considered so far have the merit of separating personal assets from those “invested” in the risk of an AIs'

activity - the company would be liable in the first place - of guaranteeing a certain degree of operational autonomy to the artificial agents, of being able to prescind from the plurality of members and of formally placing the AI agent at the company's domicile.

On the other hand, the picture of AIs being housed in legal entities of this kind is not free from criticism. Concerns may arise as to the forms of interaction between AIs and human members of the company, even within the confines of a purpose-driven estate. Both if they cooperate permanently and, *a fortiori*, if they cease to cooperate at a later date, leaving room for single-member or memberless companies. If it is true that these corporate structures are designed to foster cognitive cooperation between artificial systems and human parties, then two issues arise, as pointed out by Bayern et al.: (1) if members opt out, it will be difficult, if not impossible, to modify the operational agreement on which the hosted AIs (ergo its algorithm) is based in order to accommodate possible changes in circumstances; (2) what role would be reserved in the legal entity for developers, given that they can have a great influence on the way the AIs will perform? (Bayern, et al. 2017, 159).

This would suggest that a more ongoing and binding involvement of human parties, in the board form, is to be preferred. However, while in some circumstances creating a corporate legal platform may seem reasonable – e.g. for AIs that are used by multiple associate individuals – this method is not adaptable to every situation. Assuming that the diffusion of such technologies will be pervasive in society, to the point that everyone will have access to individual ownership or use of an AI systems (e.g. self-driving vehicles), the creation of a dedicated company for each of them, with its own statute, risks being a costly, inefficient and fiscally demanding regulatory choice for single subjects. This appears to be especially true for AIs which are not intended to engage in entrepreneurial activities.

In other words, a company housing AIs can be a feasible solution only where the associative purpose and the relations between members are well-defined. And although one can abstractly use single-member or memberless LLCs, the outcome of applying these schemes on a large scale is still uncertain.

5. AIs as individual persons in law

The legal statuses under civil law hitherto examined do not seem to cover some of the issues associated with the novelty of AI technologies, like machine learning and deep neural networks.

Treating an AIs as a mere cognitive tool implies adopting different liability law schemes and, consequently, different ways of managing externalities, costs and opportunities. Some of these schemes succeed in dealing with situations where the role and potential blame of the human parties involved can be traced, while they risk being unsuitable for those situations where the damage is caused by an incident of the artificial agent's autonomous agency (and from distributed knowledge it is driven by).

Treating these artificial systems as part of juridical persons is a stimulating proposal, but one that still leaves room for reflection on the convenience of assigning individual personality status - i.e. not consolidated by corporate-type relationships - to these same AI systems.

In this section I shall try to argue that the attribution of self-standing legal personality (*uti singulus*) on an AIs may represent, under certain circumstances, an optimal model for risks and costs distribution, at least for those situations not satisfactorily tackled otherwise.

5.1. The limits of conventional liability models

As long as we consider an AIs as a mere cognitive tool, we have at hand both fault-based and non-fault-based (or even no-fault at all) liability patterns, holding liable the human actor involved on each occasion. These patterns differ in the weight they attach to the psychological element and in the exonerating evidences.

The former, which are mostly negligence liabilities, hold liable the person who has failed to observe the appropriate standard of care. This type of approach is likely to work where the duty of care, i.e. the conduct required, is discoverable and virtually ascertainable in court. Judicial definition of duty of care, when it comes to new technologies, is a key passage, but one that requires adequate knowledge of the technology involved. Also judicial assessment of the standard of care helps to calculate the inherent risk and the degree of acceptance of a given technology (risk knowledge) (Zech 2021). Where the failure is traceable back the accountable party – be it the programmer, the manufacturer,

the data developer or the operator – negligence based model incentivises all the human agents involved, including third parties, to behave as diligently as possible.

However, setting the standard of care effectively tends to be compromised by the operational and knowledge extraction paradigm of modern AIs and the way in which artificial agents and humans interact. Autonomous skills frequently correspond to a lack of control on the part of the human user, and this is crucial if AIs are involved in legally relevant activities, like entering into a contract or negotiating the terms of an agreement. If there are multiple courses of action an artificial agent can follow in order to achieve the objective at hand, then it will be often impossible to predict precisely which of these paths it will follow under changing circumstances. This is particularly evident for more sophisticated AIs that also exhibit a kind of ‘creativity’ in the decision-making process. One of the most famous is AlphaGo, the software that has learned to play the game of Go in a professional manner (Silver, et al. 2016). AlphaGo’s programmers knew how the underpinning neural network worked and also knew the data the system was trained on, yet at several points they could not predict the exact move the software would make or the exact game tactics it would follow (and it defeated the reigning champion, by the way).

It seems then that two of the cornerstones of personal responsibility are missing: a sufficient degree of awareness of what happens (the epistemic condition) and control over the actions performed (the control condition) (Coeckelbergh 2020). It is worth remarking that these conditions are poorly met both by users and by the developers themselves, who are often unable to predict the specific course of action an AI will take.

If a standard of care (and development risks) cannot really be relied upon or placed due to information gaps of this kind, then the profitability of negligence liability decreases. This leads to an accountability issue: either humans are held accountable for actions they may not have intended to take, nor could reasonably have prevented, or there is a liability gap, as nobody can be held accountable. The result is an increase in legal uncertainty which can affect the production, therefore the improvement, of AI technologies and access to justice for the injured party (Zech 2021).

The other group of liability law solutions above considered, conversely, go in the direction of providing legal certainty in the first

place, i.e. strict liabilities like those for products or ultra-hazardous activities. The party held strictly liable, aware of the embedded risk, may internalise the costs to be incurred or pass them on to others (e.g. final consumers) and consequently plan their businesses with greater certainty. The burden of proof would be significantly eased for the victim, who would have a better chance of being awarded compensation. In addition, the party exposed to liability will find it convenient to work on improving technology and enhancing safety and reliability standards, thus ultimately benefiting the general interest.

Yet, this approach is also not without drawbacks: although it seems to be suitable for some high-risk AIs – not widely used – it may be economically burdensome to do so for all AI devices, i.e. to systematically offload the costs of accidents onto a single party; especially when the costs of for compensation are unexpectedly high. If the latter is the manufacturer (or the coder), such an approach might disincentivise production, and discourage consumers to behave diligently as long as they know that someone else will always be held liable; while if the unique liable party is the end user, maybe indirectly, strict liability might act as a brake on purchase and obviously would not put the right incentives for manufacturers to innovate and improve their products. As a matter of fact, tort law rules play a role in the size of the industry and the price of goods: negligence liability systems are associated with lower costs in supplying goods, since the producer can escape liability by meeting legal standards of care, whereas strict liability systems are associated with higher costs. As a consequence, industry will be larger with negligence-based liability, because of the effect on the demand curve of a lower price of goods (Cooter 1991, 23). A downturn for a relatively young industry such as AI could have detrimental effects including, for instance, a lack of social trust in new technology and a loss of innovation.

On top of that, relative strict liability schemes still requires proof of causation, which would not be always straightforward for AIs “[...] due to the increasing connectivity, situations may arise where distant contributors (e.g. by supplying faulty data) cannot be detected. In cases where an immediate causal person acting in accordance with duties of care can be determined (e.g. the operator of a damaging hardware component), while a more distant contributor acting in breach of duty is not held liable, strict liability may result in an incentive for lower levels of care for the distant parties” (Zech 2021, 7). Therefore, the typical

effect of the strict liability scheme, i.e. the internalisation of the costs (of accidents), is undermined as it can be too demanding for the victim to identify the perpetrator(s) and the causal link in order to file a lawsuit. If the victim is disincentivised to claim compensation, while perpetrators benefit from areas of impunity, then tort liability rules do not function properly.

In some situations, therefore, negligence and strict liability schemes risks being a sub-optimal solutions for the social interest, unable to guarantee either the victims' right to compensation or to introduce the right incentives to encourage technological innovation. On the other hand, company law solutions also seem to be hardly compatible with a scenario of a high diffusion of AI systems (as already stressed at Section 9).

5.2. The function of legal personhood on AIs

At this point, the question remains whether conferring autonomous legal personality on AIs is an efficient regulatory strategy for dealing with opaque situations (marked as 'accidents'). Also, possibly, intertwined with the solutions observed so far.

The status of legal personality links a centre of interests to bundles of rights and duties – or just one of them – and is triggered by multiple and divergent incidents. It is frequently claimed that personality can be conceived as a cluster concept, as we will elaborate in the second part of this thesis. The difference between the status of *things* and that of *persons* – which is common to most legal systems – is primarily anchored in the moral intuition that persons cannot be objects of use, they have own interests, exhibit agency and their conducts, with the resulting effects, can be attributed to them; on the contrary things are object of use, they do not possess will, interests, competence or accountability (nor then legal standing) (Benson 2002). To some extent, AIs challenge this distinction since they are equipped with an epistemic and practical authority over their behaviour that is partially autonomous from their owner or operator; to some extent they are more like animals than things.

However, given the functionalist view that I am subscribing to, what matters on a purely regulatory level of analysis is mainly the way personality status allocates legal positions, i.e. whether a new point of imputation distributes efficiently risks, costs and opportunities. In the

case of AI, a great emphasis will be put on externalities of AIs' activity and failures (namely, accidents).

Then the question to be asked is: are there any practical advantages in creating a new and detached personality status to serve an artificial agent? This question has always been addressed by the doctrinal reflection on corporate legal personality, which saw in this artifice – or 'fictions' according to a no longer successful view (D. Gindis 2009) – a means of simplifying the underlying relationships, both between members and with external parties (e.g. creditors), through an appropriate legal addressee as a replacement for a complex network of contracts.

The company operates as an autonomous legal entity, and its legal personality remains unaffected by events involving the numerous and evolving legal personalities of its managers, workers or shareholders. The convergence of interests, or powers, in a single centre of imputation of legal positions raises regulatory and coordinative questions as to how they are to be exercised and protected.

Corporate personality is then mainly justified by the need to minimise transaction costs, efficiently coordinate several individuals and the opportunity to differentiate liability, risk and tax treatment (Coase 1937; Meckling and Jensen 1983; Freund 2000); not least, by simplifying the underlying relationships, both between members and with external parties (e.g. creditors), through an appropriate legal addressee rather than through a network of contracts. The legal personality of corporations, and thus the partitioning of assets to secure both members and creditors, has historically produced stimuli for economic growth (Hansmann, Kraakman and Squire 2005; Huff 2003).

I believe that similar considerations can be made for legal personhood of AIs; there are various potential opportunities to be borne in mind: legal simplification, transparency, efficient distribution of externalities produced by AIs, better protection of injured parties and the benefit to society of having such artificial agents being more self-sufficient in activity. The expectation is also to stimulate the growth of such an innovative and profitable industry.

We know that the externalities produced by an AIs can be addressed through the rules of tort law without necessarily depriving the AIs themselves of the legal status of 'things'; However, we also know that traditional solutions do not lead to efficient results in all circumstances, e.g. unpredictable accidents (f3).

Personhood consists in the ownership of (one or many) subjective legal positions and liability is one of them; the hypothesis to be considered here is whether the comparative advantage of treating IAs as persons is to hold them directly liable. Unlike in criminal law, though, sanctions for civil wrongdoings are typically pecuniary and accordingly the status of personality under civil law can be sensibly claimed only if there is an asset assigned to the artificial agent. Thus, the hypothesis to be contemplated assumes that an AI has its own assets over which victims and creditors will be able to exercise their right to compensation in the event of a civil wrongdoing. Having an artificial agent's own assets would allow the limitation of liability and asset segregation – therefore risk mitigation – that frequently informs the organisation of economic and collective activities in the form of juridical persons. In this way, depending on how the status of legal personality is designed, it will be possible to spread internalisation costs over several stakeholders while segregating their individual assets, with the associated rewards in terms of economic incentives for production and innovation.⁶³

As far as the technical-legal aspect is concerned, having a single point of imputation of legal positions is both a means of simplification and a guarantee for the victims too (whether third parties or the users themselves). In fact, legal personality on AIs – and so their legal standing – would make it less costly to identify the accountable party in view of the litigation, rather than the requirements of negligence or strict liability rules on causation and standard of care. Much less difficulties would be encountered in proving the causal link because the ‘owner’ of the action – whether intentional or unintentional – would be the AIs in the first place. Instead, the benchmark for assessing diligence would depend more on the technological state of the art – perhaps affirmed through certification standards – rather than on the conduct of the human party. Also from this perspective, it would perhaps make sense to combine compulsory insurance mechanisms where the state of the art criterion does not help to protect the victim appropriately.

⁶³ It can be argued, in opposition, that asset segregation does not stimulate innovation since producers do not find it as uneconomical to introduce goods with very high standards of reliability to the market as they would under a strict liability regime; I will come back to this point when addressing the issue of counter-arguments to the personality of AIs.

Finally, if the rules on corporate insolvency were followed, separate assets could ensure that the victims of AIs would have priority over the personal creditors of shareholders.

To benefit from the status of legal personality, AIs should then be entered in a dedicated register, which presupposes that they meet the requirements of a specific certification and registration procedure (Sartor 2002). Perhaps according to the arrangements suggested by the European Commission regarding conformity assessment and registration of stand-alone AIs in an EU database.⁶⁴ While registration might identify the ‘domicile’ of the system and disclose the amount of the patrimonial warranty, certification will report its technical features, e.g. degree of autonomy, self-learning capacity, error rate, sources and methods of knowledge extraction and so forth. It goes without saying that these procedures would improve the transparency and reliability of the AIs introduced on the market.

From the operational point of view, it seems reasonable to presume that relieving the humans involved of an overly burdensome liability would facilitate AIs performing their functions more frequently in an unsupervised manner. And this would be promoted not only by indirect effects of portioning liability, but also because legal personality can entail holding active positions of power, as the capacity to act. At this stage, it makes sense to ask whether asset independence serving the capacity to act of an AI agent has advantages or whether, on the contrary, it should be viewed with concern. It goes without saying that we cannot make an absolute statement here, and much will ultimately depend on the rate of accuracy and success of AI technologies. Whereas for some systems it will be possible to take advantage of operational self-sufficiency – with the associated ability to enter into legally relevant interactions (e.g. high-frequency trading agents, but also some self-driving vehicles) – for others it remains preferable to tie behaviour to the supervision of a human being (e.g. autonomous weapons). The social acceptability of self-sufficient and free-acting artificial agents – on a par with other legal entities – will vary as the degree of precision and reliability varies.

To conclude on this point, it is not unnecessary to reiterate that I am not suggesting the indiscriminate appropriateness of this regulatory

⁶⁴ European Commission, Brussels, 21.4.2021, Com(2021) 206 Final, 2021/0106(Cod), Proposal For A Regulation Of The European Parliament And Of The Council. Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.

option, but that I am mainly limiting it to the combination of two factors: the unpredictability of the conduct of AIs – that set of causes that provoke what I have previously called 'accidents' (f3, see sec. 2, Ch. 2) – and the social impact, i.e. those cases in which AIs involve security, fundamental rights and values – situations that the European Commission associates with the high-risk AIs (r2, see sec. 4.2, Ch.1).

5.3. The AIs' legal personality

The resolution of the European Commission with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)) is well known in the relevant literature. What is of interest for our analysis is point 59 (f) which, in the section on liability provisions, suggests considering the implications of an 'electronic personality' status: “[...] creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently”.

To take up the European Parliament's advice, one should have an approximate idea of the type of personality status being tested. While the function that legal personality might serve is abstractly definable (even if granted on the basis of cognitive artefacts), it is much more difficult to define the exact content of this status as well as the type of economic involvement of the human subjects concerned. And the comparison with the alternative scenario, i.e. the legal status of 'thing', depends largely on how the status of person is designed.

Personality does not necessarily have to assign a fixed set of legal positions, e.g. rights, duties, powers, immunities, liabilities or liberties (Hohfeld 1913). The set of subjective positions may vary according to the entity considered, and thus we will have cases in which the personality attributes only rights - e.g. natural and animal entities (under certain legal systems) - or, rarely, only obligations - e.g. compensatory funds.

Before even designing the specific 'dress' of civil law personality of the AIs - which is, incidentally, a rather too ambitious task to be fully realised except as a theoretical experiment - it is necessary to clear up some ambiguities: one can speak of personality both in instrumental terms and in constitutional terms.

Constitutional personhood denotes that legal status which guarantees the protection of individual rights and fundamental freedoms of human beings. Rights and freedoms that are precisely recognised by the constitutions of legal systems. These are mainly, but not exclusively, prerogatives associated with being human, e.g. the right to bodily integrity, freedom of thought, freedom of religious belief, freedom of speech, etc. If AIs will gradually acquire human attributes, then it is reasonable to expect that the progressive acquisition of such attributes will be followed by a progressive equating to the legal status of human beings; but it would probably make little sense to imagine that current AIs, as well as the next generation, would enjoy such personality status. Both for factors related to cognitive competence and for those related to the degree of social integration of such artefacts (Andreotta 2021).

Yet, as Lawrence Solum observes, Yet, as Lawrence Solum observes, the discourse turns out to be more about justification because the same prerogatives can be justified both functionally and constitutionally. Indeed, the conferral of many constitutional rights and freedoms on AIs can be justified on a functional basis: “Granting AIs freedom of speech might have the best consequences for humans, because this action would promote the production of useful information. But assuming a different justification for the freedom of speech can make the issue more complex. If we assume that the justification for freedom of speech is to protect the autonomy of speakers, for example, then we must answer the question whether AIs can be autonomous” (Solum 1992, 1257).

According to the approach taken so far, I will appeal mainly to instrumentalist/functionalist justifications and less to essentialists/constitutional ones. Although a clear break between these dimensions is not necessarily desirable (as I will argue in the second part of the thesis).

In brief, it is certainly possible to explore some constitutional justifications (Solum 1992; Chopra and White 2010) – and I shall refer to few arguments of this kind – but in the perspective of regulatory policy it seems a minor investigation for two reasons at least: (1) there is nothing to prevent us from assuming that personhood is a mere legal instrument, and can therefore be invoked to serve objectives that are independent of AIs as such, like cost allocation and pursuit of social opportunities (e.g. economic growth); (2) it is not always the case that legal personhood is driven solely by distinctively human attitudes;

attributes like conscience, empathy, feeling or moral sense may not be necessary or sufficient conditions for personality (Matthias 2008).

5.3.1. The content of AIs' legal personality

Against this background, what bundles of rights and duties could make up the (functional) personality status of an AIs? In some cases personhood consists in the entitlement to a series of legal positions that must be enforced through or on the initiative of other legal persons. A classic example is that of corporations, (unborn) children or persons with impaired capacity, who find in their principal (representatives, parents or guardians) those subjects protecting their interests and implement their personality prerogatives (e.g. suing or being sued). Animals, natural entities or idols might be added to the list in the future.⁶⁵ Other legal persons, on the other hand – i.e. adults of sound mind – deliberately dispose of their own bundle of legal positions. These two forms of are also known as dependent and independent legal personality: "A dependent legal person can only act through the agency of another legal person in exercising some or all of its legal rights. An independent legal person is not subject to any such restriction and is said to be *sui juris*" (Chopra and White 2011, 159). The status of the independent legal person generally coincides with constitutionally guaranteed protection. This discrimination certainly involves certain constitutive attributes of the entities under consideration, but this does not prevent us from asking whether a dependent or independent personality status is more appropriate for AIs.

The scenario of an AIs dependent personality is theoretically more comfortable, as it would definitely not be the first time that statuses of this kind have been attributed to artefacts (e.g. corporations, ships, idols, patrimonies etc.), and seems quite in continuity with the idea of holding AIs as parts of collective legal entities. Such a status might consist of passive legal positions alone: the right to bodily integrity, to not be the object of someone else's property, to own property, to receive by

⁶⁵Actually, there are legal systems in which such statuses have already been recognised, to the extent that the term environmental and animal personhood is used in this regard. Some examples are New Zealand and India, which have recognised legal personality for rivers. See, in particular, (Brunet 2019). See, more broadly on the topic (Pietrzykowski, Personhood Beyond Humanism Animals, Chimeras, Autonomous Agents and the Law 2018).

inheritance or donation, to participate in a succession, to be entitled to compensation for damages, to be represented in a legal transaction or in court, and so forth. These legal stances constitute what is generally known as *capacity for rights* or *passive legal capacity*: “Passive capacity is properly identified as an entity’s capability in law to be the beneficiary of some legal provision or provisions, in the sense that these provisions are interpreted as aiming at protecting such an entity from some harm or at advancing some interest or another of that entity” (MacCormick, *Institutions of Law, An Essay in Legal Theory* 2007, 86). In few words, those who hold this type of personality are generally only right-bearers.

Nevertheless, dependent personality can also consist of active legal positions; for example, although corporations can only act through representatives, they have a series of active powers such as entering into a contract or participating in a corporate merger, i.e. they have the *capacity to act*. Generally speaking, capacity to act can be defined as the power to produce (valid) legal effects both in one's own legal sphere and in that of others and, eventually, to incur the corresponding liabilities. Holders of this type of personality are not only right-bearers but also duty-bearers.

Moreover, it cannot be ruled out that a status of personality consists of active positions alone, as was historically the case with slaves (or animals in criminal trials) who, as dependent persons, had only active legal positions, i.e. responsibilities, without enjoying any of the other protection or claims (Kurki, 2019).

The set of rights and duties that make up the capacity to act is very broad, also changing according to the specific entity and can be split into two sub-groups, transitional capacity and liability capacity (MacCormick, *Institutions of Law, An Essay in Legal Theory* 2007, 89).

Transitional capacity describes the group of powers to enter into or create legally salient relationships through one's own action and deliberation, producing valid legal effects able to bind oneself and others. The network of relationships and legal effects may be manifold: contracting, transferring property, suing, donating, making a valid will, contracting marriage, voting or being voted, serving as a trustee, registering domicile, and so on; not to mention all the prerogatives of collective legal persons (e.g. merging). It clearly does not make much sense to attribute some of these positions to an AI, as is the case with voting rights. It might, though, make sense that an AIs, for instance, would not only be the assignee of an asset, but would have the ability to

dispose of that money – more or less freely – for transactions and investments.

Special faculties, on the other hand, which follow from special (and acquired) qualifications - e.g. the authority to prosecute of the public prosecutor – can be defined as *competences* and are a further subgroup of the capacity to act. Transitional capacity makes it possible to participate in such interactions by producing valid legal effects, i.e. by creating positions of claim for judicial enforcement in cases of violation or non-fulfilment (MacCormick, *Institutions of Law, An Essay in Legal Theory* 2007, 93). Liability capacity, on the other hand, is the susceptibility to legal imputation for civil (or criminal) offences and wrongdoings. It makes sense to distinguish it from mere capacity to act since subjects capable of producing valid legal effects may not be held liable for any unlawful consequences. One example is that of minors, including those under the age of 14, who are frequently authorised to take legal action for small purchases or to freely dispose of the assets at their disposal even though they do not enjoy full capacity to act; another recurrent example is that of politicians who, as diplomats or members of Parliament, enjoy immunities linked to the performance of their political duties (while retaining full capacity to act); finally, there may be a suspension of liability for activities that are legally relevant but produced on behalf of and in the interest of others: in these cases the acts performed by the agent – in the Italian legal system when the mandate is *with* representation – fall solely within the legal sphere, and therefore within the liability, of the principal.

Given that dependent personality status can be so articulated, it may not be necessary to confer *sui juris* status on AIs in order to fulfil some of the functions described in the previous section, i.e. to guarantee these systems a certain degree of autonomy and to handle the thorny issues of liability. In fact, it would be possible to ensure that the AIs have assets that are separate from those involved, just as limited liability companies remain legal entities dependent on their members.

How far these assets will be able to ensure the optimal allocation of resources, given the different positions at stake, will depend on the contributions that will flow into it and on how the personality of the machine itself will be designed. One solution could be to bring into the AIs assets financial contributions from all human parties involved in development, service provision, production and use of such products; almost in the form of a compulsory insurance or compensation fund,

thus reproducing a no-fault system (i.e. absolute strict liability) (Ziemianin 2021).⁶⁶

In contrast to simple insurances or funds, however, the AIs could also actively dispose of (part of) the assets conferred on it, and possibly generate new gains with which to increase the security of any creditors or even just amortise management costs, e.g. lowering the participation fee. Revenues may come either from the AI agent's own activity and services - think of commercial agents and financial agents, but also diagnosis systems and creative agents (whether of software or works of art) - or through investments complementary to the main employment. A share of these gains could be reserved for the stakeholders and another directly for the AIs, with the option of reinvesting them or simply keeping them as collateral. Wealth accumulated by AIs could then also be taxed (Oberson 2019; Huettinger and Boyd 2019; Atkinson 2019).

Much of the actual financial autonomy will depend on the powers granted and the degree of independence of the resulting personality: i.e. in the case of a dependent legal personality, it is to be expected that earnings will be put primarily at the service of the stakeholders who will set the guidelines for management and financial prudence. Alternatively, letting the imagination run wild, it could be the case that, after raising additional funds, AI agents themselves buy their own insurance: “If the AI could insure, at a reasonable cost, against the risk that it would be found liable for breaching the duty to exercise reasonable care, then functionally the AI would be able to assume both the duty and the corresponding liability” (Solum 1992, 1245).

At any rate, it is precisely in the deliberative autonomy – albeit partial – on the management of both their activities and their economic resources that the discrepancy between AIs housed in juridical persons and AIs as individual legal persons stands out.

This personality-insurance mechanism would bring benefits similar to those derived from the creation of juridical persons, e.g. asset

⁶⁶ Google has recently come out in favour of compulsory insurance in a report on AI governance: “Alternatively, in some circumstances (e.g., where the costs of adjudicating liability are high, and the deterrence value of individualized liability is low), governments and insurers may want to consider compulsory insurance programs. Google would support discussions with leading insurers and other stakeholders on appropriate legislative models” Perspectives on Issues in AI Governance, Google, p.28.

segregation, transparency, judicial simplification and creditors' pre-emption. In particular, the system will have to undergo registration, certification procedures and could use a digital signature tool to authenticate itself. The first procedure will serve to ensure reliance on assets, both counterparties interacting with an AIs - e.g. for commercial purposes - and active contributors will be made aware of the amount of money available to the artificial agent.

On the other hand, certification, and thus the declaration of the technical peculiarities about system operation, risk level and degree of autonomy, may help to set the insurance premium charged to the parties concerned. This assumes, of course, that there is sufficient information to establish, for example, the likelihood of failures, the extent of the damage – to set minimum and maximum covered - as well as the type of sufferers.⁶⁷

Finally, the digital signature could be used to identify the individual AIs, to contrast some tracking issues - e.g. risk of software duplication – and ultimately make the legal interaction valid.

What is more, victims will no longer have to prove either the causal link or negligence of the human actor, as it will suffice to show that the damage was caused by a source covered by this sort of insurance (Zech 2021). Therefore, it will be decisive to bring clarity to what events are covered by the AIs' asset guarantee. As I have already stressed, it is accidents due to the AIs' autonomous agency, opaque knowledge processing and the extraction of information from remote and untraceable sources that pose intriguing liability issues (Sec. 2). It has to be said that this could appear just as complicated. Indeed in the model I am advocating, which does not exclude liability of 'humans in the chain' for negligence and malpractice, it would still be necessary to show that the origin of the failure is wholly attributable to the machine; and this may not be an easy task.

If these are the main events covered by an AIs personal asset/insurance, then this requires two things on the liability side. The first is that such events should not be excessively infrequent, otherwise the economic justification for the efficient distribution of costs through a shared stake in the assets of AIs risks becoming untenable. If such events were indeed rare, then an ordinary non-fault-based liability

⁶⁷ On the utility of a mandatory insurance scheme for AIs - and specifically AIs who lead corporations – see, among others, (Armour and Eidenmüller 2020).

scheme (even absolute) might suffice. Although this point would require further reflection. The second implication is that if only such events would be covered, then it would be appropriate for such a system to be combined with fault-based liability mechanisms; patchwork of this kind are constantly being tried out by different legal systems (Wagner 2012).

A mixed discipline must then discriminate along two axes: both (1) the AIs *type*, e.g. only those showing sophisticated processing and learning capabilities will be entitled to personality, and (2) the *outcome* of an AI system's conduct – e.g. if the outcome is caused by a clear mistake in programming, manufacturing or misuse of the system it will be desirable to apply traditional liability methods (e.g. product liability or vicarious liability), so as not to encourage the concealment of malpractice or foster suboptimal equilibria of innovation and professional diligence; otherwise the direct personality/responsibility of the AIs would remain.

As far as the first aspect is concerned – the distinction between models of AIs – a three-stage approach can be adopted: ranging from the minimum level of autonomy for which the legal status is essentially that of a cognitive tool (to be associated with custodian liability maybe); the intermediate level for which some legal autonomy begins to be conferred (maybe housing the AIs in human-controlled juridical persons); and the maximum level which is associated with the proper status of individual legal personality (Dahiyat 2021, 75). A further classification criterion to be used in conjunction with mere operational autonomy could be the risk of social impact (and harms) of the AI agent: so that, for instance, robots with a high degree of autonomy but a negligible capacity to cause considerable social damages would remain treated as a cognitive tools. In this latter case, conventional liability rules would continue to apply.⁶⁸

The second operation, i.e. the distinction based on the type of AIs outcome/failure, is likely to be more complex because it is often unclear which conduct actually emerged from artificial agency and which from human mistake or recklessness. However, negligent behaviour can be attributed to either the manufacturers (or programmers) or the users,

⁶⁸ The risk criterion and the adoption of a similar system, with the difference that high levels of risk and autonomy correspond to strict liability regimes, was proposed in the expert group report on AI liability to the European Commission (2019): <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608>.

because even if the latter have much less supervisory powers than the former, it cannot be ruled out that a malfunction is due to a lack of monitoring or diligent use on their part. Where this reconstruction is not possible, direct liability of the AIs might be preferred.

These last questions raise a more general point, not limited only to liability profiles: is it consistent to confer an individual legal personality on an entity and then restrict it to certain circumstances? Of course it is, especially if one considers that the status of *dependent* personality allows certain agents to be treated as legal persons solely for certain activities and purposes (e.g. only for contracting law), and according to the characteristics of the relationship they have with the principal. It might become a less sensible solution if, instead, the status accorded is of the *independent* type, since this would mean that artificial agents have a sphere of rights that do not serve the interests of others, but their own, in which case it would be more difficult to set major limits on the pursuit of morally recognised interests (Gray, *The nature and sources of the law* 2006).

A further idea to bear in mind, which can be seen as an alternative to the mechanism of personality envisaged so far, limits the area of assets conferred on the AIs, and therefore of legal personality, not to liability for any unforeseeable damage, but only for those produced against the owners/users of the systems themselves and not against third parties. The model is that of the so-called market enterprise responsibility, i.e. while damage to third parties is borne by the owners or users, damage to the latter is borne by the producers, who place AIs on the market equipped with an asset fund. The result is always the creation of some kind of legal personality, whereby the AIs will be directly liable to a limited extent with the assets assigned to them. The advantage is that producers can know the costs of externalities in advance.⁶⁹

To conclude, the most realistic personality scenario so far remains that of AIs with *dependent* type, holding mainly duties and responsibilities and fewer rights. It seems premature to discuss conferring independent type of personality on AIs, despite the fact that some of the conditions for this recognition – as often cited in the literature – do not seem much beyond the reach of current technology. Samir Chopra and Laurence F.

⁶⁹ Here we return somewhat to the idea of compulsory insurance – of the type of ‘Guarantee Fund for Road Victims’ – with all the limitations that this solution brings with it.

White, in this sense, refer to five attributes: intellectual competence, susceptibility to legal obligations, susceptibility to punishment, contract formation, and economic and property holding capacity. It seems that many of the skills above are already owned by current AIs. Still, it is worth waiting for technical progress to make more convincing arguments on this point.

In any case, in line with the approach of this thesis on optimal allocation of risks and opportunities, the expected result of the combination of personalisation and negligence liability rules is to prevent failures due to negligence and to parcel out the costs of unforeseeable accidents so as not to discourage the production, consumption and technological innovation of AI devices.

5.4. The case against legal personality on AIs

Woodrow Barfield and Ugo Pagallo recognise three main trends in the AIs personality debate: “(1) those who believe such AI systems could or should already have the status of legal person; (2) those who are open to accept such a scenario in either the mid-term or foreseeable future; and (3) those who suggest we should never accept any of the above two scenarios” (Barfield and Pagallo 2020, 64). At this point it should be clear that I tend to converge on the second trend. This section, however, deals with the third trend, i.e. those sceptics (more or less radical) that AIs ought or can be given any form of legal personality.

Since the regulatory hypothesis of giving legal personality to AIs was mooted in the European Parliament resolution in 2017, several criticisms and concerns have been raised. In 2018, a group of experts set out quite succinctly in an ‘Open Letter to the European Commission on Artificial Intelligence and Robots’⁷⁰ the problems that, in their view, the former proposal brings with it; and subsequent European reports on AI have not taken up the suggestion of the so-called electronic personality.⁷¹ So,

⁷⁰ Open Letter to the European Commission Artificial Intelligence and Robotics, link: <http://www.robotics-openletter.eu/>.

⁷¹ See, for instance, the report from the expert group on liability and new technologies appointed by the European Commission: ‘Liability for Artificial Intelligence and other emerging digital technologies’ (2019): <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608>; and also the expert group ‘AI HLEG’ (2019):

it could be said that, in a certain sense, the thesis of the legal personality of AIs is a minority view both in the scientific and political debate.

I will briefly review what are the main arguments against such regulatory hypothesis, trying to figure out whether they are justified and, when they are, whether it is possible to rebut them. These critiques cut across both civil and criminal law personality status. It should be pointed out that there are objections of a mainly ontological nature – closer to an essentialist conception of legal personhood – and objections of an instrumental nature – invoking futility or dangerousness of this choice of legal policy or its inability to achieve the intended objectives.

Although the functionalist reading of personality that I subscribe to might exonerate me from addressing purely ontological objections, I think it still makes sense to discuss both classes of criticism, since the property-based approach is not entirely unconnected with the pragmatic grounds underlying the use of legal categories such as personhood. Yet, I believe, ontological/essentialist objections are less capable of debilitating the regulatory proposal.

5.4.1. Ontological objections

Objections of an ontological/essentialist nature are often preceded by a generic accusation of anthropomorphism. But this accusation is quite unfounded – and might be countered with the accuse of anthropocentrism – since there is no perfect correspondence between the set of legal persons and the set of natural (or moral) persons. At most, there is a one-to-one correspondence, i.e. being a natural person implies being a legal persons, but the reverse is not the case. This simple observation would suffice to dispel much of the rhetoric surrounding about the charge of anthropomorphism.

Yet, among the ontological reasons preventing AIs from being conferred personalities are generally the cognitive and moral traits of human condition. The lack of these qualities, according to some authors (Solaiman 2017; Gordon 2020), would prevent AIs from being able to effectively hold the prerogatives and duties associated with legal

<https://ec.europa.eu/digital-inglemarket/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>.

personhood. It would therefore not only be an inherent problem of what these systems *are*, but of what these systems *do* (Chesterman 2020).

The distinctive qualities of the human condition generally assumed to be presupposed by a full and effective right to legal positions are: intelligence, conscience, self-awareness, sensitivity, (the experience of) free will and moral agency. Doubts therefore arise as to whether AIs are intellectually competent enough - or have the right form of competence - to understand legal obligations and to dispose of their legal sphere; the 'missing something' argument as Solum calls it (Solum 1992, 1262).

Having personality, and thus being the addressee of a set of legal norms, tend to imply the capacity for these norms to be understood by the addressee. The lack of proper comprehension is not a problem for juridical persons since it is still the human members who determine how the collective agent will acknowledge the legal rules and behave accordingly; while it could be a major issue for AI agents.

The first point to make against this remark is that the validity of this kind of objections is limited to independent legal personality, whereas dependent personality does not appear to be affected. Beyond that, there are some aspects that seem to be underestimated when making statements about the intellectual uniqueness of human beings. First of all, as already sustained in the first chapter, the AIs are smart enough to act on the basis of reasons, trying to optimise the achievement of one or more goals, interacting with the external environment and being able to manipulate (also in a creative way) the received information as well as to implement new ones as perceived by the environment and by their own experience. In this sense, it seems possible to describe AIs as intelligent agents, capable of participating intentionally in relevant legal situations – e.g. contracts - without the ongoing supervision of the human principal.

On the other peculiarities of the human condition we still know very little: e.g. the way our conscience works, the role it plays in our self-awareness or deliberations, whether there is a free cause for our actions etc. We know, of course, that current AI systems – and likely next generations as long as this computer architecture is used - are not comparable to humans in terms of consciousness and meaning attribution (i.e. semantics) (Dreyfus 1992). But, as far as legally relevant cognitive capacity is concerned, this is not a conclusive argument against the personality of AIs.

I would opt again for a pragmatic approach: the problem is not to establish what are the general cognitive peculiarities of human beings, but which of these count for the purposes of legal practice and for the entitlement of active personality, i.e. composed of both rights and duties. I believe that the key functional attitude here is that agents take legal rules as *reasons for action*: that is, as premises for the practical reasoning which justifies or explains their behaviour (also called reason-responsiveness). AI agents acknowledging such practical (and epistemic) authority to legal norms would reveal their sensitivity to obligations.

While the protection of self-consciousness, moral maturity or (the experience of) free will by legal personality appears as a 'privilege' of some entities only - i.e. adults of sound mind - it would be difficult to make agents who do not perceive legal rules as reasons for action participate in legal interactions. Humans with such deficits are still treated as legal persons, since they maintain a moral status, but one may not see any need for it in the case of non-human entities that are not accorded any moral status (yet).

At this stage it should be stressed that although artificial agents cannot manipulate semantics, they can still engage in reason-responsive and obligation-sensitive behaviour. (Dignum 1999; Castelfranchi et al. 2000; Anderson and Anderson 2007; Boella and Van Der Torre 2007).

For this purpose, in fact, it seems sufficient for AI agents to have *instrumental rationality*, i.e. one that is useful for pursuing an interest (one's own or that of the user) by virtue of deliberations that take account of the possible consequences of one's actions and the way in which these consequences - also represented in the form of legal sanctions - may compromise the satisfaction of the interest itself (Lagioia and Sartor 2020, 440). The fact that the instrumental rationality enabling AIs to recognise and acting on legal norms does not coexist with the capacity to be morally involved in them, is not an obstacle to the holding of many of the legal positions embedded in personality.

Quoting Chopra and White: 'Work in deontological logics or logics of obligations suggests the possibility of agent architectures that use as part of their control mechanisms a set of prescribed obligations, with modalities made available to the agent under which some obligations are expressed as necessarily to be followed'. (Chopra and White,166). And AIs may have rational motives, embedded in their basic drives (as we have seen in section 2), to comply with legal obligations, since failure to do so could compromise essential drives as self-preservation or

maximisation of expected utility. This last profile also implies some form of susceptibility to punishment (as argued in relation to criminal personhood).

Some might object that being mechanically compelled to follow rules is not the same as having real sensitivity to obligations, but this is not obvious from a technical point of view: some artificial agents may sometimes be allowed to violate legal norms when the infringement results from their own deliberative process as being more beneficial (Castelfranchi, et al. 2000). As in the case of the agent who violates privacy rules for fear of catching a virus from a malware (sec. 5.3.1). This clearly poses significant problems where the violation is systematically carried out by an AIs for reasons of its own convenience; and the possibility of accidents caused in this way are at the root of some of the liability issues to be solved (also) by the conferral of legal personality. On the other hand, one might claim that the chances of conforming to virtuous standards of moral conduct are much higher in artificial agents than in natural ones. As some authors ironically suggest, while humans only tend to deontological or utilitarian ideals of moral agency, artificial agents might be totally aligned to those moral standards (Brożek and Bartosz 2019).

In any case, even if current cognitive qualities of AIs rule out their having proper moral agency – for which the requirement of consciousness and sensitivity might be strongly required, (Véliz 2021; Gibert and Martin 2021) – they seem to meet minimal requirements for participation in legal practice since they can engage in autonomous conduct that can be explained and predicted in terms of intentions and normative beliefs. My point is therefore that from a strictly attitudinal point of view AI agents already possess the necessary capacities to recognise legal norms, produce legally valid effects and to be held liable for them,⁷² but that since they do not have full moral agency their hypothetical personality status will not for a long time be comparable to that of natural persons. It follows that for these reasons, and for those related to the functional dimension of personality, the "missing something argument" is not able on its own to defeat the regulatory hypothesis under consideration.

⁷² Here we should disentangle liability and (moral) responsibility.

5.4.2. Instrumental objections

Instrumental objections mainly point to: the danger of human parties escaping responsibility, the disincentive to improve technology, dangers of letting AI agents have legal initiative and identification issues and, finally, the overall futility of applying legal personality (Bryson, Diamantis and Grant 2017; Solaiman 2017; Dahiyat 2021). I take this last criticism to have already been addressed in the preceding sections that attempt, precisely, to argue the pragmatic value of personality for AIs.

The phenomenon of evasion of legal liability is certainly not posed for the first time by AIs, but has long been the subject of debate in the field of company law with respect to the presence of separate legal entities. The limitation of investors' liability that is achieved through corporations gives rise to concerns on the part of creditors when the assets of the corporation are not able to satisfy claims for payment. In order to counter this abusive practice, the doctrine of the 'piercing of the corporate veil' has taken shape, which makes it possible to overcome the barrier posed by asset segregation and hold the company's stakeholders directly liable (Macey and Mitts 2014). The concern that the same evasive phenomenon also occurs in the case of an AIs' financial inability to repay debts – due to accidents caused or breaches of contract – is heightened by the fact that, in contrast to corporations, in AIs as autonomous legal subjects it would not even always be possible to identify the relevant stakeholders (Bryson, Diamantis and Grant 2017, 288). I am not really convinced by this objection. The attribution of personality has, among other functions, also that of making the underlying relations between the parties involved in the supply chain (and use) of AIs more transparent. Consequently, I do not find any particular obstacle to the application of the 'piercing of the corporate veil' doctrine also to insolvent AIs. Moreover, the patrimonial capacity of an AIs should be mandatory communicated to third parties - or in any case to be declared at the time of registration on a hypothetical register of 'electronic persons' - and any operation not adequately covered could be unauthorised (also by design). While all wrongdoing intentional committed by human actors through AIs - both crimes and torts - would remain covered by conventional forms of liability.

A second type of objection is that assigning personalities would not create the right system of incentives and rewards with respect to

technological innovation and thus would not promote damage prevention. I believe that this objection is fairly well-founded, since if part of the liability is shared between several parties - including AIs with their own revenues - then it can be expected that the costs of possible accidents will be less on the shoulders of those parties tasked with research and development of safer systems, i.e. manufacturers and designers. In other words, if the cost of making the product safer exceeds the expected gain from the improvement by too much, as the cost per injury remains stable, then the objective of injury prevention may fail and overall welfare suffers. However, one way to defend against this objection is empirical observation of what happens in other markets. There does not seem to be a vicious cycle like the one just described, for example, in the automotive market simply because: (a) consumers continue to have an interest in buying safer and safer cars and (b) safety devices are required by law to be certified as compliant. The economic interests of the parties involved does not remain invariant as the failure rate of the systems changes: more reliable AIs are more successful on the market, and produce fewer economic losses, both in terms of fees to be paid for damages and in terms of service interruptions, for all parties involved.

There are then criticisms concerning the identification of AIs (as software agents of course) and the dangers of letting them have legal initiative. I believe that both are challenges that can be addressed through technical as well as legal stratagems.

Difficulties in identifying individual AI agents may undermine personality assignment, since these agents have no physical location and may duplicate or reproduce themselves in multiple software. In these cases it becomes impossible to understand which agent is accountable for an action and how it interacts with its environment. In any case, a registration and certification system can curb such problems: “The solution may lie in using digital signatures. Anytime the agent uses a signature, to sign a legally relevant action, it is uniquely identified. That is to say, although software can be copied, keys or signatures are protected in a key-vault, and are protected against copying” (Dahiyat 2021, 60)

On the other hand, the fact that AI agents are free initiators of legal interactions is not a problem in itself, if they have been trained to comply with legal rules, but the fact that, by self-learning skills (ML), they can change the way they function, may be. However, the self-transformation

of AIs can be tolerated up to a certain point beyond which the system itself can be induced to stop working, perhaps by invalidating the digital signature system that gives official status to the legal actions it performs (Dahiyat 2021, 61).

In any case, problems linked to the identification of AI agents and the way in which they must be registered are likely to persist until widespread technological patterns are consolidated on the market. This aspect is part of the more general, albeit trivial, question of the feasibility of an 'electronic personality', i.e. that it is not just a choice of convenience of the legislator or jurisprudence, but must also wait for certain environmental, infrastructural and technological conditions to come about.

Finally, I believe that one of the greatest concrete difficulties that may be encountered by the regulatory hypothesis of conferring legal personality on AI systems concerns the quantification of the capital with which to equip them. In fact, in order for the calculation of this capital to be reliable, it is assumed that it is possible to outline the operational risk, the externalities, of an IA. However, this is often not really feasible precisely because of the unpredictability of the operation and decisions of some AIs. The danger is that either too little capital is placed on it - which would undermine the rights of creditors (victims or otherwise) - or too much - which would make the regulatory option uneconomic.

Tackling this problem can only mean defining more precisely the level of operational risk that each AI system entails. It seems to me that this is a possibility encouraged by the European AI Act itself, which, when it proposes a classification of the risk of AIs mainly based on the type of rights and interests involved, also indicates that it is necessary to specify the risk analysis to the specific application contexts.

This kind of information gap is common whenever a new good is introduced on the market and contributes to the problem of coordination of legal policy choices between legislators that I will discuss in the next section.

6. Why do we need theoretical analysis of legal personality: the coordination problem

So far I have tried to emphasise that the regulatory option of the 'electronic personality' is still a game in town. However, I would like to conclude by emphasising a further issue that needs to be clarified in

order to assess the feasibility of this hypothesis. In fact, legal personality attribution to AIs seems destined to face some sort of coordination problem. Even in the scenario where there is widespread (political) agreement about the personality of AIs, different rules will be required depending on the technical peculiarities of AI systems, and maybe on the sectors in which they are used. The definition of AIs itself – think of one of those provided in the first chapter - still struggles to find general applicability and does not duly discriminate according to certain technical peculiarities that may be of great importance from the regulatory perspective.

Then, it should be established whether the resulting personality covers all legal areas or just some of them (e.g., only civil law); personality is an aggregate status made up of different positions that are not necessarily joint, as demonstrated by the fact that two entities with this relational property can hold very different bundles of rights and duties. And the forms of protection or enhancement of interests enforced by personality status may vary according to the legal system of reference with its specific disciplines.

Sources of complexity are partly due to the fact that legal personality is triggered by diversified conditions and consists of heterogeneous bundles of legal positions capable of creating as many personal statuses as there are legal fields⁷³ and partly related to the sociotechnical implications, due to the application context we are discussing. Complexities of this kind often prevent us from taking a clear position on regulatory policies, and disrupt coordination mechanisms when decisions of this kind have to be made in heterogeneous legal contexts.

In other words, there would be a sort of coordination problem, both within and among legal systems, that can be broken down into three types of asymmetries:

(a1) *Technological*. Each AI can perform very different tasks with various kinds of risk, autonomy, and social impact; the technological peculiarities of each of these tasks (as concerns both software and mechanics) may call for specifically tailored regulations.

⁷³ Specific reflections on the nature of the concept/property of legal personality will be devoted to the conceptual analysis in the second part of the thesis.

(a2) *Intra-system*. Each body of law (e.g., criminal law and company law) may assign to the notion of personhood slightly different meanings entailing differentiated rules and outcomes; these mismatches often lead to disagreement and misunderstanding among lawyers and legal scholars.

(a3) *Inter-system*. Each country's legal system may have a different set of rules concerning legal personhood, generally exercising exclusive jurisdiction over these rules; this is a matter of relevance in contexts like that of Europe.

In a nutshell, it is easily predictable that in such a regulatory scenario, as many notions of legal personality would arise as there are types of artificial systems and fields of application, e.g. financial algorithms, autonomous weapons, personal care and assistance systems, for private or public transport, and so on. Of course, legal rules and institutions are designed to set more or less stable equilibria for situations, such as this one, characterised by uncertainty and lack of coordination. Among the functions that legal rules can pursue, and that may prove useful here, are also that of standardization and reduction of information costs.

Anyway, in this case I believe the regulatory intervention to be hindered by the characteristics of (the concept of) legal personality and from the strong technological instability. Not leaving out the possibility that some of the technology asymmetries may be independent of the state of the art, and thus persist in the long run. The first profile cannot be tackled just by abolishing the differences between personality types and personality conceptions assimilated by various legal systems. On the other hand, the jurist can do little about the latter profile, apart from encouraging standardization. From a more theoretical perspective, what I think jurists can do is to reduce uncertainty with new classifications, conceptual tools and doctrinal categories in anticipation of future breakthroughs on extra-legal floors. This is how I interpret the philosophical reflection that follows in the second part.

To conclude, although there are some functional reasons for granting legal personality to AIs, it would be worth clarifying what use should be made of the concept of personality and how these coordination problems could be solved. I think that the 'noisy' and uncertain debate on AIs personality shows that a wider theoretical reflection on legal personality is needed also to address issues pertaining to the conferral

and contours of legal personality: some conceptual order is required even before entering into value judgments. This does not exclude that similar points could be raised for other entities.

In the following analysis, my claim will be that jurists can generally rely on metalegal and intermediate concepts in managing complexities of this kind with a view to finding equilibria in the legislative debate; this could also work for the legal personality ascribable to AIs.

7. Conclusions

AIs can behave independently from their human creators and users, displaying a kind of intentional agency which raises issues of responsibility for their behaviour (and for any unintended consequences). These issues come specifically into play considering that some of the conditions for attributing responsibility for the behaviour of autonomous AI systems — e.g. the conditions of ‘control’ and ‘awareness’— may no longer be fulfilled by the individuals using those systems. Thus, conferring legal personality on the latter may be justified by the need to fill the resulting liability gap.

Some of the available legal measures for regulating the use of AIs and their autonomous agency both under criminal and civil law have been discussed. Arguments were then sought to express the inadequacy of conventional models of liability under certain situations. Namely, in case of tortious liability, either a fault is established for negligent supervision by custodians or strict liability is placed on owners or custodians regardless of any failures of them in controlling the AIs. Yet, as stressed, in some cases custodians cannot be blamed for negligent supervision, given the autonomy of AIs and the technical inability to foresee their specific behaviour. On the other hand, strict liability risks being unfair, or at least too severe, for the custodian, especially where the damage that needs to be redressed is unexpectedly high. Placing strict liability on the shoulders of producers or programmers, on the other hand, risks discouraging production and innovation without ensuring victims’ protection and compensation. Even greater problems may arise in criminal law, as the cognitive preconditions for criminal liability may not be met by the humans using AI systems.

I have therefore claimed that two possible approaches, aligned by the creation of a separate asset with which creditors can be secured, can be embraced to address these problems: (1) the first may consist in the

creation of a company or other legal entity (collective or not) with personality that has as its constitutive purpose precisely the activity of an AIs; (2) give personality directly to AIs, empowering them to hold legal positions on their own, act autonomously, generating legally valid acts (e.g. contracts) and producing legal effects on third parties, maybe with complete financial autonomy.

Each of these two approaches has its limitations and have not been presented as definitive solutions. Specifically, I have argued that personhood may be an option for all those cases where: an AI system does not require human review or oversight to function properly; certain socio-technical conditions are met (e.g. risk profile or application context); its harmful or unlawful behaviour is not due to misuse of the system, nor is it due to auditable or foreseeable defects in the system (e.g. that existed at the time of its introduction to the market), but rather results from the way it functions, according to its design or interaction with inference models, data, human parts, or other devices (i.e. those cases that have been labelled "accidents"). Given these limitations, it has been cautiously suggested that personality-based schemes would still benefit from combining with negligence liability rules.

**PART TWO – A THEORETICAL
FRAMEWORK OF LEGAL
PERSONALITY**

I. The Philosophical Inquiry of Legal Personality

1. Introduction: philosophical tools

So far, legal personality has been mainly observed as an instrument for the distribution of legal positions responding to pragmatic and economic rationales. In particular, among the pragmatic reasons that may justify conferring personality to an entity there is the expected utility of the better distribution of risks, benefits and externalities as realised by virtue of allocating legal positions to a separate point of imputation. Similarly to what happens for collective legal persons, I claimed that, under certain circumstances, the allocation of legal positions (especially liabilities) directly to an AIs is desirable when it provides a better way of managing social costs and opportunities arising from these technologies – as well as providing the right incentives – than conventional legal schemes.

The previous chapter aimed to provide a reconstruction of the liability issues that arise when artificial intelligence entities act beyond a certain degree of autonomy and predictability. To this end, it was suggested that the creation of a new point of imputation for legal positions – i.e. personality in Kelsenian words – can be functional in distributing risks and costs arising from liability over a plurality of subjects. This distribution may turn to be, under specific circumstances, the optimal one.

However, this functional reading gives only a partial picture of what legal personality is. In fact, the motives inspiring the assignment of this status are not always strictly functional and, in any case, what makes the distribution of legal positions efficient may depend on further reasons intertwined with the way legal practice operates. In short, even the functional reading, in order to work correctly, must assume that specific properties and values are relevant to legal personality.

In this part of the thesis, legal personality will be depicted in a comprehensive manner, that is, by means of a metaphysical survey (combined with a conceptual map) showing the relationships between the positive law norms governing the status and the deeper set of reasons justifying them. Legal personality will be represented as an institutional legal concept consisting of a triad of rules (both constitutive and regulative), according to an approach of social ontology, sometimes labelled as Neo-institutionalism, which finds in the work of Neil MacCormick one of the most known elaborations for the theory of law.

More specifically, I shall provide a model of the metaphysical structure of legal personality, as neutral as possible with respect to particular philosophical conceptions of its nature, e.g. monist or pluralist, legalist or realist (see below), or with respect to philosophical views about, e.g., the human nature. The aim of this metaphysical enquiry is to reconstruct the building blocks and the background sources of legal personality, which is seen as a socio-institutional kind/fact, through a scheme potentially applicable to other legal kinds and facts.⁷⁴ However, the arguments I defend in the first part of this thesis are suitable for a functionalist metaphysical anchorage of legal personality.

Some notions of analytic metaphysics – e.g. *grounding* – will be used to illustrate the forms of constitutive and explicative relations that link certain facts and properties to legal kinds. These relations occur at different ontological layers: it will be argued that legal personality is *constituted* (grounded) both (a) in a set of conditions (e.g. facts and properties) which are fixed by legal rules and (b) in a set of conditions that operate in the background and that, rather than being fixed by, shape legal rules. To these main ontological layers more can potentially

⁷⁴ I will explain what is a social (or institutional) fact/kind later. For the time being, I just want to point out that legal kinds and facts are a sub-category of social (or institutional) facts and kinds and therefore I will often use only the first expression.

be added, even intermediate ones (which I will try to do), as long as they do not prove redundant.

I shall then present a multi-layered ontology of legal concepts, which consists of an institutional, an intermediate and a meta-institutional layer. Each layer describes sets of facts, values and properties relevant for the personality status. The content of the institutional layer is fixed by legal rules, while the other two layers tend to be extra-legal, in a sense that we will elaborate on later. The intermediate level includes concepts that, like mid-level terms in morality, can help to generate reflective equilibria in the legislative dialogue. Thus facilitating theoretical agreements in circumstances complicated by indeterminacy, as seems to be the case also for AIs.

The point is that the distinction between metaphysical layers of facts determining legal personality, to which correspond peculiar relations of existential (non-causal) dependence, seems consistent with the architecture outlined in the conceptual analysis: namely, the set (a) of conditions which are fixed by legal rules can be framed as the institutional layer, while the set (b) of conditions that operate in the background can be framed as the *meta-institutional* layer.

I will also consider the opportunity of describing institutional and meta-institutional layers through detached metaphysical relations. In particular, attention will be paid to the innovative model proposed by Brian Epstein to design the structure of social reality, and which makes use of three different metaphysical relations: grounding, framing and anchoring (also GAF model). I will suggest the possibility of applying this model, or the intuitions behind it, to explain the metaphysical structure, the building blocks, of legal personality.

I shall also stress that the philosophical analysis carried out in the second part of this thesis does not have a purely theoretical purpose, but is also driven by a concrete need. Indeed, the 'noise' surrounding the regulatory hypothesis of AI personality requires clarification on: (1) what are the origins, presuppositions and modes of existence of legal personality; (2) whether alternative strategies and conceptual tools are available by which information and rules that tend to be associated with personality in such a way as to facilitate convergence among different views in the juridical dialogue.

1.1. A functional and ontological account of legal personality

Some might object that the functionalist reading of personality is incompatible with the ontological account of this Second part. However, this objection risks confusing separate but consistent enquiries. As will be seen, in fact, the ontological analysis of legal personality aims at reconstructing the relations of existential dependence between the legal kind/fact of personality and the kinds/facts that constitute it as such (I will elaborate more on the distinction between causality and metaphysical grounding later). Investigating the metaphysical grounds means identifying what makes the case for legal personality, thus shedding light on the constituents and also on the assumptions of legal rules. Yet, this enquiry is neither analogous to, nor incompatible with, the enquiry into the justifications and pragmatic implications of treating AIs as persons of law. Here possible facts (i.e., legal arrangements) are explored. This latter is much closer to a kind of causal explanation, while the former is into metaphysical explanation.

Somehow Brian Epstein also recognises this difference, distinguishing between the project of 'model building' and 'grounding inquiry'. In the Part one of this thesis, I have adopted a 'model building' approach: "Much of the practice of tax law, for instance, is concerned with exploring different possible structures, to see whether they satisfy the conditions for being taxable. When Verizon bought out the stake held by Vodafone, was that a taxable transaction? What if they had structured the transaction in a different way? This is an investigation of possible facts, and the application of frame principles to evaluate whether various other facts would ground the possible legal fact *The buyout is taxable*?" (Epstein 2015, 100).

I believe that both accounts are needed, for theoretical and pragmatic reasons, which in the case of AIs intertwine: (1) to explore the effects of legal compositions in possible worlds, in our case the possible world in which AIs have personality, it is necessary to have an exhaustive picture of what law is; (2) to address the coordination problem, we should examine the ontological presuppositions and the various modes of existence of legal personality with a view to constructing balances in the debate about law policy choices (more on this later).

Although the conceptual and metaphysical framework I introduce aims to be neutral with respect to the various theories of legal personality, the 'model building' operation I perform in the first part obviously leads me to lean towards a certain conception of the metaphysical structure of personality. In particular, I consider that what

holds together the constituent elements of legal personality for AIs, building legal rules in a certain way, is a functional reason.

2. Conceptual analysis and real definition

I shall address different, though interlinked, set of issues: what is meant by the concept of 'legal personality'? What other concepts and semantic fields are connected with legal personality? And also: what it takes for something to be a person in law? What is it like to be a legal person? What is it that makes the status of a person in law persist over time?

These are, of course, questions that answer different philosophical enquiries: (a) the first three fall under that properly metaphysical enquiry, also called 'real definition', committed to defining the mind-independent properties of things, thus considered as real objects of the world; (b) the second instead observes the same objects as linguistic or mental constructs, and is thus committed to an analysis of the theoretical role played by terms and connections with other concepts, as well as underlying conceptual and semantic practices.

In this part of the thesis I wish to provide both a metaphysical and conceptual framework of legal personality that will hopefully also be useful to the regulator. Conceptual analysis, on the one hand, is focused on the ordinary meaning of concepts, their logical structure, uses, referents and connections with other meanings/concepts.

Real definition, on the other hand, is focused on the features of the object itself; in our case of the legal property of being endowed with personality status: what it is for an entity to be a legal person or, in other words, what necessary and/or sufficient conditions an entity must meet to qualify for this legal status, what effects this status implies and so on (many of these aspects have already been outlined in the previous section) (Rosen 2015).⁷⁵

Although they are different philosophical enterprises, somehow starting from incompatible assumptions – e.g., the first is typically semantically internalist, while the second is externalist – they can lead to

⁷⁵ As Gideon Rosen puts it, exemplifying the concept of courage: “*what it is* for a person to be courageous — to identify that in which the courage of the courageous person consists — by specifying non-trivial necessary and sufficient conditions for courage somehow grounded in the nature of courage itself” in (Rosen, Real Definition 2015, 189).

convergent conclusions (also for the sake of policy-making). In particular, I believe that conceptual analysis, although not a necessary precondition, can be functional to the metaphysical explanation of legal personality. Though with different methods and objectives, conceptual analysis and metaphysical explanation can be used to identify the set of properties that make the case that something falls under a given (real) kind or concept. A legal kind/concept in our case.

2.1 Conceptual Analysis

What is meant by conceptual analysis needs to be clarified. I partly subscribe to that group of theories, often labelled the 'Canberra Plan', which traditionally sees conceptual analysis as consisting of: (1) collecting the network of claims, principles and terms – sometimes generically called *platitudes* – that explicate the role a certain concept-term implicitly plays in our theories; and (2) locating the thing in the world that plays the theoretical role, e.g. a descriptive property (Nolan 2009).⁷⁶

First of all, what are 'platitudes' and how do they contribute to informing a concept? Broadly speaking platitudes are claims about the subject matter mirroring the ordinary use of a concept-term, but the precise answer about their nature and role may depend on the general view one has of concepts. At first glance, one might say that since concepts are linguistic devices, their nature depends on how one conceives of language itself. This is not necessarily wrong, but one should avoid identifying concepts with language: they can be seen as representational mental states, as specific cognitive skills, or as abstract objects like Fregean senses (Margolis Laurence 2021).

Subscribing to one conception or the other is not straightforward, not least given the peculiar types of concepts concerned in this thesis, namely legal concepts. Indeed, legal concepts, embedded in legal norms, hold together a large amount of information and are suitable for covering divergent situations (especially if they are nodal ones). For this

⁷⁶ This two-step procedure of the Canberra Plan was outlined by Daniel Nolan. Anyway I do not subscribe entirely the Canberra Plan approach, but more the Jackson's variant.

reason concepts like property, marriage, and personality are seen as *cluster concepts*, even though I disagree with this idea (as we will see later).

Against this background, I shall consider two different stances on legal concepts. The first, more sympathetic to the ‘Canberra planners’ – e.g. Frank P. Ramsey, Rudolf Carnap and David Lewis (Braddon-Mitchell 2009) – is a reductionist and eliminativist reading, which has been echoed in the legal domain by Alf Ross (Ross 1957). Legal concepts in this view, specifically ‘intermediate’ ones, would be like the theoretical terms of scientific theories according to Ramsey, i.e., definable simply by the role that what they refer to plays within a given theory. As a result, the concept-term might be eliminated and replaced, as Ramsey sees it, by existentially quantified variables collected in sentences explicating the relevant information of the platitudes; (Nolan 2009, 268)⁷⁷ or by links between factual preconditions and deontic conclusions, as suggested by Ross. Either way, the task of conceptual analysis would be to make *explicit* the theoretical roles concepts play within certain theories.

According to this reductionist reading, and the inferentialist variant that can follow (we will see that), concepts like legal personality would simply consist of links between set of use conditions and juridical effects; knowing what is meant by the term of legal personality would be equivalent to knowing what it is like to be a legal person.

However, this approach does not seem exhaustive: conceptual analysis can additionally display the deeper network of meanings, philosophical commitments and social beliefs in which a concept is embedded, i.e. further concepts it seems to imply or presuppose (Himma 2015). Capturing this deeper network allows for better targeting of discussions on broad issues – e.g. what are we talking about when we talk about justice? – to frame theoretical disagreements in a clearer way – e.g. what are the assumptions of our conception of justice? – and to address complex questions about the extension of a concept-word to certain cases – e.g. would we say this case is covered by our concept of justice?

Thinking of platitudes in this different way leads us to the second view of (legal) concepts for which their meaning is instead *compositional*, i.e. it results from the articulation of terms associated to the conceptual category and from the specification of the connections the same has with other terms. From this perspective it is the lexical components

⁷⁷ The original theory can be found in (Ramsey 1931).

associated with concepts that compose their meaning rather than exclusively the theoretical or inferential links between the sentences in which the lexical components are involved.

For the time being, it is sufficient to note that this second view on concepts enables relevant information to be organised in conceptual frameworks, also called *ontologies*. This thesis will elaborate an ontology of legal concepts – inspired by Ota Weinberger and Neil MacCormick's neo-institutional theory of law – that explains information pertinent to personality.

However, these two visions of concepts are not incompatible, as we shall see. What is more, they are both functional to the main achievement of conceptual analysis according to the Canberra Plan project, more specifically in Frank Jackson's theory (which differs from the standard model due to the absence of the notion of platitudes).⁷⁸

In Jackson's version conceptual analysis is intended to reveal whether the way of describing something with a certain vocabulary is made true by a description of the same thing with a more *fundamental* vocabulary (Jackson 2000, 61). This means partitioning possible cases (aka 'worlds') into cases where facts covered by the description obtain and cases where these facts do not obtain. As, for example, the conventional conceptual analysis that describes the concept of 'knowledge' in the (more fundamental) terms of 'true', 'justified' and 'belief'; thus providing an account of which cases fall under this description of the concept of 'knowledge' (Chisholm 1957; Ayer 1956). Thus depicted, concepts are representational devices that play a cognitive function in the way we divide the world into states of affairs, and can do so both when thought of as theoretical nodes and as components of a larger ontology.

While not immune to well-founded criticism, much of which complains of inapplicability to all types of concepts,⁷⁹ what is significant about Jackson's vision is the way conceptual analysis combines with, or

⁷⁸ Conceptual analysis à la Jackson is not even equivalent to traditional so called Oxford-style analysis, since there is no claim to express analytical truths or necessary and sufficient conditions. The task of conceptual analysis according to Jackson is rather to find fallible information through (folk) intuitions about which cases are covered by a certain term.

⁷⁹ Some criticisms of the use of Jackson's theory in ethics are given in, (Laskowski and Finlay 2017); criticism of a different nature is raised in in (Plunkett, Expressivism, representation, and the nature of conceptual analysis 2011).

is instrumental to, metaphysical inquiry, i.e. to figure out what the world is like (in a modest fashion).

2.1.1. Conceptual Jurisprudence

In this thesis the task of analysing legal concepts, in Kenneth E. Himma's words, is intended to: “[...] explicate the content of each concept and locate them among a general conceptual framework that guides both our linguistic practices regarding the relevant concept-words and our legal practices themselves” (Himma 2015, 3). Conceptual jurisprudence is the theoretical project that addresses topics of legal philosophy – e.g. what is law, where it originates, what is validity etc. – through the analysis of legal systems' core concepts. According to someone, as typically the advocates of standard conceptual analysis, an in-depth examination of the content of these core legal concepts, and of their interrelations, may serve to reveal in some way the proper nature of law. Such as, the set of facts and properties that, when instantiated, *constitute* something as a legal fact or kind. I mention ‘facts’ and ‘kinds’ because of sympathy with institutionalism (shown below), but put another way, conceptual analysis could be said to explore what makes something a *norm* or a *legal system*: “The instantiation of the relevant properties constitutes a system/norm as one of law in exactly the same sense that being unmarried constitutes a man as a bachelor. Any institutional system of norms instantiating the appropriate properties is, for that reason, a legal system; any system that does not is, for that reason, not a legal system. Similarly, any norm instantiating the appropriate properties is, for that reason, a law in some legal system; any norm not instantiating the appropriate properties is, for that reason, not a law in some legal system” (Himma 2015, 20).

Thus, conceptual analysis can serve to discriminate facts and norms that are law from those that are not. Indeed, part of the task of conceptual analysis is to produce a framework of concepts that include both legal and non-legal ones, and to explain the function the latter play in determining the content of the former and in legal practice in general. Thus, for instance, what role does a given concept of ‘morality’ – but it could also apply to ‘rationality’ – play in determining the content of legal norms and how can it determine their validity? Obviously, different views on the nature of law and its building blocks can lead to different conceptual architectures, with different interrelations, and vice versa.

One of the best-known legal conceptual analyses is certainly that provided by Herbert L. A. Hart in 'The Concept of Law' (1961). Through his analysis, Hart not only reconstructs the ordinary meaning of the word law, but also aims to describe the nature of law in terms of ontologically prioritised and foundational elements, i.e., social facts and conventions.⁸⁰

I shall attempt to make a conceptual jurisprudence exercise, albeit limited to a single legal concept, personality, by embedding it into a framework that includes non-legal concepts which appear to determine it, while maintaining an ontological independence from the legal domain.

2.2. Towards a conceptual and metaphysical framework

But if concepts are mainly representational devices serving cognitive functions and produced by linguistic practices, why should their analysis tell us anything about the nature of the matter they describe? This is a quite debated issue.

An articulate response is advanced, again, by Frank Jackson. According to the author, any 'serious' metaphysics should not address its canonical issues – i.e., *what is* and *what is like* – just by drawing up simple catalogues of the existing, but should rather provide an account of the subject matter in terms of basic, or more fundamental, notions (Jackson 2000, 4). Indeed, our comprehension of reality would not benefit from an indefinite list of all the ingredients that can be considered as more or less relevant for a subject matter, but from setting a progressively smaller, and limited, number of ingredients compared to the initial ones.⁸¹

For Jackson, serious metaphysics would then be by nature concerned with the question of where this limit should be set and, accordingly, it would be constantly involved in a theoretical challenge addressable only

⁸⁰ But this is not uncontroversial: see, for instance, the critique formulated by Ronald Dworkin in (Dworkin 1986).

⁸¹ This view lends itself to a number of objections that question the very necessity of using conceptual analysis to carry out what is ultimately a 'smooth reduction'. However, Jackson replies, even where conceptual analysis seems dispensable it turns out to be presupposed in defining the use and meaning that, for example, scientists attribute to certain phenomena before reducing them to more basic phenomena (so he gives the example of the concept of temperature in thermodynamic theory scientifically reduced to the kinetic energy of molecules). In (Jackson 2000, 57-59).

by means of conceptual analysis, i.e. the so called *location problem*: where to locate less fundamental features in a description of the world expressed in a more fundamental language (like that of physics); or whether to eliminate less fundamental features, perhaps because they are only putative, i.e. not justified by such a description of the world. So, for instance, although being distinct, the property of *density* is not an additional feature of the world to the properties of *volume* and *mass*, and, consequently, we can count on a more basic account of the world relying on the latter two notions alone, because they already contain (*entail* in Jackson's thesis) that of density (Jackson 2000, 4).

If it is true that metaphysics concerns when matters described in one vocabulary are made true by matters described in another (more fundamental), then this requires establishing, by means of conceptual analysis, whether it is correct to describe matters in terms of the various vocabularies and thus which possible cases fall under these descriptions (Jackson 2000, 41). For Jackson, the strong connection between conceptual analysis and metaphysics finds expression in the so called *a priori* entailment thesis: a complete microphysical description of the world entails *a priori* – i.e., with justification independent of experience (or: conceptually) – ordinary macroscopic truths (Chalmers and Jackson 2001). Yet, the assumption that the availability of a concept allows one to infer metaphysical truths is highly controversial (Diaz-Leon 2011).

Still, does providing a conceptual analysis help to ultimately *explain* the essence of the world it describes? Jackson's answer is negative. Here we observe an important discontinuity with the traditional project of the Canberra Plan. In particular, while the exponents of the latter project attribute to conceptual analysis the task of discovering the essence of the world, with no concern about linguistic and social practices (*immodest role*), for Jackson: “Conceptual analysis is not being given a role in determining the fundamental nature of our world; it is, rather, being given a central role in determining what to say in less fundamental terms given an account of the world stated in more fundamental terms” (*modest role*) (Jackson 2000, 44).

This means that in the modest approach, conceptual truths express only the way in which a community commits itself, through certain social and linguistic practices, to a certain vision of the metaphysical world structure; in the immodest approach, instead, conceptual truths express truths about the world independently of human practices (Himma 2015, 27).

The tension between conceptual analysis and metaphysical explanation can be addressed in two ways. In a weak way, by advocating a modest conceptual analysis. This means that we believe conceptual analysis can help to think, state and frame problems in a clearer way. Possibly, it contributes to discovering the actual occurrence of something in the world given that to ascertain whether something occurs (or not) in world we first need to state under what circumstances we would agree that it does (Kingsbury and McKeown-Green 2009).

But then one can also respond in a stronger way, invoking the immodest approach. Indeed, as we shall see, legal reality is an institutional reality anchored (also) in mental phenomena, that is, in conceptual phenomena. As the way we think about legal reality affects the nature of it, conceptual analysis is also a metaphysical explanation. This does not mean that immodest conceptual analysis matches the metaphysical explanation of any reality, or of all institutional realities. It simply seems to do so for legal reality, that is, for its kinds and facts, as will be seen.

Against this background, I shall deal first with the conceptual architecture of legal personality and then with the ontological building blocks of such legal kind. I shall then claim that conceptual analysis and metaphysical inquiry converge on similar conclusions: the metaphysical structure of (the kind/fact of) legal personality appear consistent with an updated version of the neo-institutional theory of legal concepts.⁸²

The study of the ontology of the legal personality, and of its constitutive elements, invokes the use of the diagnostic tools of analytic metaphysics. The perspective of the study of the ontology of legal kinds and facts to which I shall refer is that of ‘Social Ontology’, concerned with discovering the nature and origin of social entities. Indeed, similarly to the method of conceptual analysis described above, social ontology seeks to provide an account of the social world through links between different facts or kinds, i.e. by identifying *in virtue of* which facts, and in what way, social facts (or kinds) occur. I will use *metaphysical grounding* as the analytical tool to describe existential priority relations among kinds and facts.⁸³

Grounding is a non-causal constitutive relationship between facts (or kinds) such as the more fundamental fact grounds the less fundamental

⁸² I will address the distinction between social facts and kinds later.

⁸³ For an overview on the topic see, among others, (Correia and Schnieder 2012).

one. So, just to give few examples, the fact that I got a speeding ticket may have been *caused* by the fact that I was late for work, but it is *constituted* (grounded) by the fact that the car was travelling on the motorway at a speed of 30 km/h over the limit. Or, also, the fact that there is a teachers' strike may be *caused* by pay cuts, but it is *constituted* by the fact that teachers picket outside the school. In other words, the grounded fact, e.g. the strike, occurs *in virtue of* the grounding fact, e.g. that teachers picket outside the school instead of teaching; this is the metaphysical reason for a given kind or fact, e.g. the strike.

Some properties of the grounding relationship are that it is asymmetrical, synchronic, transitive, irreflexive and non-monotonic; it will be seen what is meant and how these features work.

In general, the grounding relation assumes that reality (social and non-social) is hierarchically ordered and that questions about its structure can be metaphysically *explained* by existential dependency relations between facts or types. The explanatory purpose of the grounding relation is also marked by its asymmetrical nature: the grounded fact exists *because* the grounding fact exists and not vice versa, implying that it is metaphysically necessary that if the latter is obtained then the former is also obtained. When this explanation is the closest version to the observed phenomenon, and explanatory gaps are excluded, then the *explanans* is said to be *constitutive* of the *explanandum*.

In line with other works on social ontology (Epstein, *The Ant Trap: Rebuilding the Foundations of the Social Sciences* 2015), I will make use of metaphysical grounding to provide an account of the building blocks of social/institutional kinds and facts and, namely, of a peculiar class: legal ones. Inspired by this inquiry are questions such as: what is it that constitutes, makes the case for, legal personality, what makes X a legal person, why certain conditions rather than others determine it, and others.

As mentioned, this will be proved to be consistent with the findings of the conceptual analysis also because of the suitability with an institutional reading of law, with specific reference to Ota Weinberger's and Neil MacCormick's neo-institutional theory of law (MacCormick and Weinberger 1986). The reason for this is that the peculiarities of the metaphysical structure of social – and thus legal – kinds and facts call for sharpening the analytical tools, perhaps by introducing different forms of grounding or discriminating between grounding and other metaphysical dependency relations (e.g. anchoring). Roughly, I claim

that it does not seem possible to account for the metaphysical structure of legal personality solely by means of first-order grounding links because facts lying in the ontological background affect the mode of existence of legal kinds, and they do it in different ways. These backstage facts pertain to the noninstitutional layers in the conceptual framework (meta-institutional concepts).

My aim is not to speculate on which of these constituent facts represents the ultimate metaphysical essence of the legal personality, e.g. which is the predominant idea of humanity or the strongest legal-systemic reason, but only to provide an ecumenical model of the metaphysical structure of the legal personality, and perhaps applicable to other legal types. I will have space to argue this idea in the next chapters.

II. A conceptual framework of legal personality

1. Introduction

In this chapter, I carry out a conceptual analysis of legal personality. The aim is to identify the main meanings conveyed by this legal label and the semantic domains to which they belong.

Two possible approaches to legal concepts, the inferentialist and the ontological, will be considered. The merits and demerits of these two approaches will be exposed, but the emphasis will be on how they can reinforce each other. In keeping with MacCormick's neo-institutionalist perspective, the 'ontological approach' to legal concepts will be subscribed to, also with a view to proposing a conceptual framework, seeking to update the traditional neo-institutionalist reading.

The resulting conceptual architecture will consist of at least two ontological layers: (a) institutional legal concepts, whose content depends on the rules of a certain (legal) institution, here the concept of legal personality; (b) meta-institutional legal concepts, whose content depends on the broader (social) practice of which the constitutive rules of an institution are only an instance, and which are presupposed by institutional concepts, e.g. the concept of intentionality. In the end, I will argue that an intermediate ontological level can be identified by adapting the abstractness of the meta-institutional layer to the characteristics of juridical practice, as formalised in operational legal rules.

2. Law's conceptual texture

Legal concepts channel a large amount of information. Yet, some claim that as the law is reducible to norms and principles, the role of concepts is on the whole negligible in the law ontology. This view is sometimes called *normativism* (von der Pfordten 2009). One version of this semantic reductionism, i.e. inferentialism, will be specifically addressed below (section 6).

From the inferentialist perspective, perhaps normativism in general, the law is described in its operational dimension. Still, work on concepts is an integral part of the work of legal officials and lawmakers. After all, the law is full of references to concepts of different natures, some of which are strictly technical, others are ordinary, and still others are somewhere in between, not having a fully legal pedigree; the latter class of concepts includes legal personality. In particular, it seems to me that the debate around its contours is a clear example of the extent to which law, in addition to establishing use conditions and normative effects, is engaged in an extralegal conceptual analysis of what a 'person' is and how this status is to be differentiated from that of, say, 'things' or 'animals'. In other words, law, with its proper tools such as rules and principles, is not self-sufficient in assigning semantic content to such linguistic entities. It follows that there is a dedicated space for the conceptual analysis of law – e.g. conceptual jurisprudence – which is not confined to linking bundles of rights, duties and other legal provisions.

As I stated in the introduction, I do not take real definition to be interchangeable with conceptual analysis. This, in terms of the ontological debate on concepts, moves me away from realist theories according to which concepts are non-representational properties of real entities (whether natural or social) (Carnap 1950; Peacocke 1992). Nor do I subscribe to reductionist theories of concepts like nominalism, for which concepts are flattened onto representational linguistic units such as words and/or their connections.

An intermediate position, based on a representational theory of the mind (RTM), will be favoured: concepts are systems of internal mental representations with a language-like syntax and compositional semantics. In short, according to this view – which is a quite default position in cognitive science (Margolis and Laurence 2021) – the syntax and content of concepts are determined by “the syntax and the content

of its constituents” (Fodor 1998, 94). I will return to compositionality later (sec. 7).

This reading of concepts is sometimes labelled as conceptualism: “If conceptualism is accepted, concepts are not to be understood as independent things or facts. It is sufficient to assume mental processes of differentiating qualities as conceptual. Concepts are only reifications of these mental processes. In the same way, that joy is a reification of being happy, a concept is a reification of understanding

a property. This means ‘no concept without understanding’ ” (von der Pfordten 2009, 24).

In the following sections, the concept of legal personality will be analysed and classified according to three axes: (1) application properties (fuzziness and generality); (2) meaning (inferentialism and compositionality); and semantic referentiality (eliminativist and institutionalist).

3. Persons in the legal tradition

Let me first provide some historical coordinates of the concept of ‘person’ in the legal thought, without claiming to offer an exhaustive account, but mainly with the aim of showing the transformation of its area of extension, and also the profiles of continuity with past legal traditions. Further historical exploration will be conducted in the next chapter (sect. 6) by discussing the main theories of personality and their legal and philosophical roots.

The term Latin term ‘*persōna*’ has an Etruscan root (*phersu*), probably in turn derived from the Greek (πρόσωπον or *prósōpon*): all these terms referred in one way or another to the concept of ‘mask’ or ‘theatrical character’. As a matter of fact, the term ‘*persona*’ was used to refer to the theatrical mask that Roman and Greek actors wore on stage with the somatic features of the character to be interpreted and with the purpose of amplifying the volume of the voice.⁸⁴ However, semantic evolution, marked by a series of metaphorical shifts (e.g. synecdoche), has reshaped its meaning and reference. Over time, in fact, the ‘person’ ceased to indicate only the mask, but rather the theatrical role itself and, later, coincided with the individual wearing the mask (Keeton 1930, 117).

⁸⁴ For the etymology of the word ‘persona’ you I looked up the Italian Etymological Dictionary Online, at <https://www.etimo.it/>.

Ultimately, the term began to be applied to the juridical sphere of individuals, i.e. the legal status, as a dividing line with things, animals and other entities.

Historiography suggests that one of the first formalisations of the division between 'things' (*res*) and 'persons' (*personae*) – together with 'actions' (*actiones*) – is that made by the jurisconsult Gaius in his *Institutiones* (Poste 1904). This division acquired a general organisational value in the jurist's treatment of Roman civil law, and still appears to influence most legal systems today.⁸⁵ As early as Gaius' work, the term 'person' was to describe the legal status of human beings, groups and collectives (e.g. *collegia* and *universitas*). Yet, there was no really an explanation of its distinctive juridical meaning. Rather, it seems that the qualification of person appeared as a *genus* to which all human individuals belonged, albeit with different *species* depending on their social and economic status. So, the capacity to act was not implied by the juridical personality in itself, as it could change according to the specific category or instance of personality of concern – e.g. *liber*, *servus*, *uxor* etc. – and found full realisation only in the so-called *persona sui iuris*.

The term 'things' was also mainly used in its common sense, even though in practice it also covered rights and duties – conceived as incorporeal things, the so-called *res incorporales* – and some animate beings, such as animals, slaves and women (Trahan 2008).

It must be pointed out that the extensive use of the term 'person', both in the *Institutiones* and in the *Corpus Iuris Civilis*, has led to misunderstandings about its actual area of coverage. Indeed, personality was not guaranteed, at least in its entirety, to all individuals; an emblematic case is that of slaves who, in some circumstances, had some of the persons' prerogatives – for instance, acting on behalf of the master or entering into quasi-marital relationships – while in many other circumstances they were legally equated with *res*, e.g. being the object of their master's vindicatory actions (they were *personae alieni iuris*) (Watson 1967). In any case, it would be wrong to think that slaves had a proper legal personality, e.g. there was no recognition of property rights, especially if compared to the modern conception (Buckland 2010).

The same holds true for women who, although abstractly part of the 'persons', were in fact completely excluded from Roman public law, i.e. without political, property and testamentary rights or other rights that

⁸⁵ For an overview on Gaius *Institutes* see, among others, (Stein 1999).

could be claimed in court. As a matter of fact, women were essentially equated with *res*: until marriage, they remained under the control of their father and later under that of their husband, and the only right they could exercise was that of inheriting from their husband in the event of his death (Couch 1894).

The theoretical scheme of the law of persons laid down by Gaius stood for a long time, and the first significant new contributions came only from the natural law theorists of the 16th century, like Hugues Doneau and Huig de Groot, also known as Grotius. These authors were among the first to give a more technical and less vague meaning to the concept of person in law, reconceptualizing its role in legal theory and beginning to envisage independence between the notion of human being and person in law: “Grotius added little to the stock of existing ideas, but what little he did add proved to be important: ‘persons,’ he wrote, are those who ‘have rights to things’. Though Grotius himself did not say as much, this attribute of persons clearly implies—indeed, presupposes—another, namely, that persons ‘can’ have such rights, in other words, have the ‘capacity’ to receive or acquire them” (Trahan 2008, 12). These insights were later cultivated, and enriched, by Gottfried Leibniz, who, in his attempt to systematise private law, treats the notion of ‘person’ in conjunction with that of ‘rights’; he is one of the first philosophers to conflate being a person with being a bearer of rights and duties (Kurki 2019, 38).

In the late 18th century, partly as a result of the European codification process, the notion of a person began to identify a universal legal subject, i.e. an entity endowed with general legal capacity regardless of socioeconomic condition. Yet the first theories that started to conceive of the legal person as the subject of rights and duties, according to an approach very close to modern conceptions of personality, were those of the legal positivists of the early 19th century, in particular, those of Anton Thibaut and Friedrich Carl von Savigny. In the reflections of these two authors, one can glimpse the first clear separation between the natural and legal dimensions of personality, and therefore the creation of a detached juridical space, i.e. a point where the set of legal positions connected to an individual or collective interests could converge.

As John R. Trahan claims: “This manner of (re-) defining person marks an important shift – indeed, a reversal– in thinking about ‘personality’. Whereas in earlier times ‘being a person’ was thought to be logically prior to and to be the cause of ‘having legal capacity,’ hereafter

'having legal capacity' will be thought to be logically prior to and to be the cause of 'being a person'. Second, the modern theory establishes a new 'umbrella' category into which the various non-natural persons (*collegia*, corporations, etc.) can be conveniently placed, namely, 'moral' (in the sense of 'psychological') or 'juridical' person" (Trahan 2008, 15). As is well known, in Savigny's theory, this happens through the expedient of the notion of legal fiction.

In the 20th century, one of the doctrinal turning points that have had the greatest impact on the contemporary legal conception of the person, and on substantive law, is that of Hans Kelsen. The Austrian legal positivist is credited with having inaugurated a doctrinal process of progressive dematerialisation of legal personality, which in the end becomes a mere point of the imputation of rights and duties. There is thus an inversion in the traditional way of understanding legal personality: according to Kelsen, personality *does not create* the set of subjective legal positions but is instead simply the "unit of account" (MacCormick 1988, 372) of the legal order, i.e. it is created by the intersection of specific bundles of legal rules and consists only of rights and duties. It is the law that personifies legal rules and not the rules that refer to the behaviour of individuality because it has personality: "The person as a holder of obligations and rights is not something that is different from the obligations and rights, as whose holder the person is presented—just as a tree which is said to have a trunk, branches, and blossoms, is not a substance different from the trunk, branches, and blossoms, but merely the totality of these elements. The physical or juristic person who 'has' obligations and rights as their holder, is these obligations and rights—a complex of legal obligations and rights whose totality is expressed figuratively in the concept of 'person.' 'Person' is merely the personification of this totality" (Kelsen 1945, 95).

Kelsen's work on personality is therefore consistent with his general vocation to discover the foundations of pure legal science. However, this view, which has influenced and still influences a large part of the legal and philosophical debate, has not been free from criticism of various kinds, many of which complained of an excess of formalism.

3.1. Post-Kelsenian developments and the contemporary debate

Since Kelsen, although not formalised in an organic theory of personality, Herbert L. A. Hart's contribution to the debate deserves

attention (Hart 1984). As is well known, Hart's view is largely influenced by Wittgenstein's philosophy of language, fostering an explanation of legal terms in the light of the legal rules and conventions that give them their meaning. According to this reading, the content of the concept of 'person' is expressed only by the network of existential conditions and uses under a legal system, and not by any deep metaphysical meaning. So, dealing with legal personality would not require jurists to make factual assessments concerning essential properties or values, but it would require them to understand the legal settings it is embedded by.

One of the attempts to update the Kelsenian theory of personality, while maintaining some of its basic insights, is that of Neil MacCormick. The neo-institutionalist view of personality advocated by MacCormick, and which is partly endorsed in this thesis, aims to reconcile the strictly legalistic dimension with the extra-judicial one, e.g. in order to understand how legal categories affect social conduct: "We think it possible to accept some of the great central insights of the Pure Theory of Law but to re-express them in an institutional theory which restores a sociological awareness to legal discourse and which makes plain the meaningful quality of legal categories for social agents" (MacCormick 1988, 374). Accordingly, MacCormick follows a kind of property-based model, which gives much weight to certain capacities – e.g., continuity of consciousness over time, intentionality, sentience etc. – and differentiates between active and passive aspects of the grounds of personality, triggering different legal forms of personality.

The second half of the twentieth century is characterised by a decline in interest in legal personality, probably also due to the political achievements that have led to the disappearance of some of the most problematic issues, e.g. the end of slavery and the progressive universalisation of the status of person (Kurki 2019, 52). This loss of interest is accompanied by a conversion in the terms of the discussion, which appears much more sensitive to ethical and substantive issues.

Indeed, anti-formalist critiques and substantive readings of the personality in law have been revived by the contemporary debate on human rights, bioethics and biolaw. In particular, on the first front, the notion of person has acquired a bridging role between citizenship rights recognised within the State and universally recognised human rights: the personhood condition, essential to the enjoyment of individual human rights, becomes a prerequisite for the enjoyment of citizenship rights (Canale 2015, 11).

Another direction, still within the discussion on the protection of human rights, is the one traced by Martha Nussbaum, a well-known proponent of the capability approach, relevant to various ethical and political fields (e.g. theories of justice) (Nussbaum 2000). In a nutshell, she proposes to combine a universalistic approach to human rights with an interpretation of personhood that takes into account the various roles and social contexts in which individuals exercise their capacities and meet their basic human needs.

On the other hand, the notion of person is involved in discussions on euthanasia, abortion, genetic therapies, assisted reproduction, and other typical issues of bioethics and biolaw. The opposing bioethical positions promote on these issues different conceptions of the person, and of legal personality: typically, a more substantialist and finalist one, of religious inspiration, and a more normativist and functionalist one, of secular nature (Canale 2015, 13).

Lastly, some inputs to the debate have been brought by feminist theories that have emphasised – often in an anti-liberal key – the political dimension of the (legal) personality, at times promoting the introduction of a new paradigm of recognition, i.e. that of ‘vulnerability’ (Fineman 2008).

Against this background, the contemporary discussion is no longer focused on the doctrinal framework of legal personality, while there seems to be a broad consensus that this term covers three senses: (1) according to ordinary language, whereby ‘person’ and ‘human being’ are co-extensive concepts; (2) according to philosophical-moral language, whereby a ‘person’ can be said to be an intentional subject capable of being accountable for its actions; (3) according to strictly legalistic language, whereby a person is just a point of the imputation of legal positions (Naffine, *Who are Law’s Persons? From Cheshire Cats to Responsible Subjects* 2003).

The first perspective is a relevant basis for legal discourse and: “[...] is axiomatic in discussions of ‘human rights’ in which the fact that human beings are the ‘natural’ subjects of rights from birth onwards by the mere fact of being born human, irrespective of considerations regarding their mental or physical state, is normally taken for granted” (Gindis 2016, 506). The second perspective influences whole areas of positive law, e.g. the way competencies are distributed and shaped in contract and criminal law. Finally, the third perspective expresses the function of law to pursue pragmatic aims, e.g. systemic and economic,

through the allocation of legal positions to a formal point of imputation. This purely functional-allocative use, as noted in the first part of the thesis, underlies many reflections on the design of corporate personality (Deakin, *The Juridical Nature of the Firm* 2012). It might be precisely this latter meaning of personhood that opens the way to the law of persons for those entities that are not traditionally part of it, as I believe may be the case as far as AIs are concerned

Nowadays, the debate on legal personality is then more oriented towards other aspects: for example, towards the way such a status affects social conduct; what is self-identity and how it matters; how many subjectivities one hold (political, health, gender and so on) and how the law should take them into account; towards ascertaining whether certain entities possess the relevant properties – including new features as ‘vulnerability’ – or otherwise carry interests, both personal and social, to be protected through that status.

In this context, the line of investigation pioneered by Neil MacCormick seems promising, and I will try to integrate it with a metaphysical and conceptual apparatus that clarifies some theoretical aspects that have remained incomplete. I believe that a clarification of this nature may also help to frame the current debates on the contours of personality; here with specific reference to the case of AIs.

4. “Who is law for?”

Although the category of ‘person’, and its distinction with that of ‘thing’, is still one of the foundations of Western legal systems, it is not uncommon to have disputes about its content and its numerous, sometimes incompatible, applications. I believe that the explanation for this is multi-causal: some of the difficulties are certainly due to historical motives, that is, to the fact that the status of ‘person’ in law is the product of a long-term religious, political, and socio-economic processes. Such processes are still ongoing, as witnessed by the debate on the rights of animals, fetuses, groups, environmental bodies and artificial agents (Wise 2001; Kurki and Pietrzykowski 2017; Gellers 2021). Other problems are prompted by an unclear legal framing of ‘personality’, i.e. a poor understanding of its role and pragmatics. Finally, further misunderstandings result from the polysemous nature of the term ‘person’ and by its contamination with contexts which are not strictly linked to law but which still play a role in the legal practice.

As Ngaire Naffine emphasises in ‘The Law’s Meaning of Life’ (Naffine 2009), the personality of law is first and foremost a means for identifying the target community – for Naffine the question it answers is ‘who is law for?’ – that is, the collection of entities that bear rights and duties and that are ultimately the true recipients of legal norms. The identification of the community of reference can serve and has historically served, a variety of needs. Firstly, a practical need to establish who, in what form and with what limits can dispose of rights and privileges or be subject to duties and responsibilities. Secondly, a symbolic-rhetorical need, i.e. to attribute value and normative weight to certain entities (and principles) and not to others, e.g. things or animals. This suggests that delving into the second dimension of legal personification, and the ways in which legislators and judges carry out this process of identification, implies questioning a number of related issues that go beyond the strict legalist dynamic: “[...] when making their determinations about legal personhood, lawyers and judges often feel obliged to consider whether the being in question has the necessary intrinsic or attributed characteristics to qualify for legal being; whether it is the right kind of being to be thus legally endowed. Is it considered sufficiently intelligent? Does it feel enough pain and pleasure? Is it sacred? Does it possess intrinsic value? If it is the community which is thought to endow it with value, then is that social value sufficient for legal personification?” (Naffine 2009, 4). The earliest and most immediate manifestation of the multifaceted character of legal personality concerns human beings and the idea of humanity.

The beginning of the process of association between the person in law and the human being has a Christian matrix, especially following the Council of Chalcedon (451 A.D.), the first historical event in which an individuality combining human and divine substance was recognised in the person of Christ (Canale 2015). Following this event, the term ‘person’ began to describe individualities that are free in that they have reason. This shift in the meaning of the term ‘person’ is subsequently secularised, also thanks to the philosophical contributions of authors like Locke and even more so Hume, ending up describing mainly the uniqueness of the rational nature of the human being. Subsequently, a further development of moral-philosophical reflection specified a meaning of person, or personhood, which became particularly relevant to the legal sphere, i.e. restricted to entities capable of being accountable for their actions (Dennett 1976).

It is clear from the historical factors mentioned in the preceding section that the notion of a person in law or legal person was never simply synonymous with 'human being', 'rational individuality' or 'rational individuality' or 'responsible subject' (contrary to what moral philosophy suggests). Yet, the current configuration of Western legal systems still provides for a partial overlapping of the two spheres of meaning, at least in the sense in which the natural person is spoken of. As we have seen, for a long time this was not the case: there was no perfect correspondence between being human and being a person in law and whole categories of subjects, like women and slaves (e.g. Early Common law and Roman law) (J. F. Gardner 1987), were excluded from possessing sufficient freedoms, rights and responsibilities, etc., to be declared subjects of the legal community. Among the reasons for this, some of the most incisive have to do with a particular metaphysics of the human nature, i.e. assumptions about both the building blocks of and value conferring properties of human beings – which also are deemed relevant for the legal field – and about who actually holds them.

Metaphysical assumptions still influence the contours of legal personality, even in those contexts where there is widespread recognition of the intrinsic and absolute value of the human being, perhaps driven by the affirmation of universal human rights. In these contexts, e.g., in Western legal systems in which all individuals are virtually 'full' persons in law (*sui iuris*), different views of the human nature affect specific legal statuses – e.g., criminal law status versus medical law status (Naffine 2009, 6) – while still adhering to the value of the human being as a whole.

However, persons in law are not only human beings, i.e., natural persons, but also other entities, often called legal (or artificial) persons. Leaving aside the differences between each legal system, artificial persons are generally associations of individuals, e.g. companies and corporations, that may have various internal organizations. In some jurisdictions, even the requirement that members exist is waived and personality takes on a purely instrumental value. In these cases one can think of personhood as a linguistic-legal device that encapsulates the network of relationships, and so of the bundle of rules, between (one or) more individuals, goods and assets. In this sense, someone has

thought of the personality in law as a metaphor with a cognitive task (Galvano 2010).⁸⁶

In other jurisdictions, for instance in India, the category of persons in law stretches even further, involving non-human animals, religious places or symbols, and political entities of various kinds (e.g. village councils).

I consider it wrong to think that there are no metaphysical grounds to be investigated behind the artificial or instrumental versions of legal personality. Indeed, these cases reveal that metaphysical assumptions extend far beyond human nature and involve assumptions and beliefs about the very nature of law as practice. And they might also tell us something about the social context in which legal practice is embedded.

In other words, I believe that Naffine's valid insight that the inquiry into legal personality is actually a broader inquiry about the legal community – ‘whom the law is for’ – needs to be pushed even further. The question of legal personality concerns an even broader question about the foundations and purpose of legal practice. It reflects, in a nutshell, the essential purpose of a legal system. Our concept of legal personality tells us ‘what the law is about’.

To conclude: the theoretical point of interest in this second part of the thesis concerns the way in which the concept of legal personality reflects and holds together all the key intuitions that inform a legal system. When we talk about legal personality, as well as when we talk about other central legal notions, we indirectly bring forward a multi-layered discourse involving multiple metaphysical intuitions.

A large part of the second part of this thesis is devoted to articulating more analytically these claims, presented here in a discursive manner.

5. Personality statuses and persons in law

Generally, being a person in law means being the holder of legal positions like rights, duties, liberties, powers, and responsibilities. Each position often implies further legal effects for the holder and a chain of jural correlatives for others. As Wesley N. Hohfeld argues in his theory of rights, subjective legal positions cannot be understood in isolation. To understand them well, one must consider them as an integral part of the network of logical relations implicitly established with opposite and

⁸⁶ On the metaphor of personification in general, see (Lackoff and Johnson 2003).

correlative jural positions (Hohfeld 1913). Hence, for example, the atomic (advantageous) position of ‘right’ has ‘non-right’ as its opposite and ‘duty’ as its correlative, or, the position of ‘power’ has ‘disability’ as its opposite and ‘liability’ as its correlative. From this intuitive picture it becomes clear how intricate the number of relationships and effects in a centre of legal personality can be.

As a result, we will find various legal personality statuses, which can be differentiated according to their ‘thickness’ and to the areas of law concerned; and of course in relation to the type of composition (e.g. individual or collective).

The status of legal personality will be more or less thick depending on the number (and weight) of legal positions held by a given entity. More precisely, we can specify as follows the minimal concept of a person in law, i.e., the *thinnest* concept of personality: an entity has personality under the law if there is at least one rule conferring a right or a duty on the entity (or the possibility to acquire a right/duty by triggering the required operational fact).⁸⁷ This thinnest concept of legal personality can be expanded into different personality statuses, each being conferred on a different kind of entity and having distinct legal consequences (Naffine 2009, 46).

A legal system may then have different *thicker* concept of personality, i.e., different personality statuses that, for different types of entities, trigger different bundles of legal positions. In other words, an entity possesses thick personality if it is addressed by a complex package of rights and duties. A personality status may confer such rights and duties directly or specify the conditions for acquiring them.

Typically, in modern legal systems the broadest personality status is conferred on all human beings: they enjoy all fundamental rights and can acquire all kinds of legal positions established by general laws (though some limitation may be established under particular conditions, e.g., incapacity).

More limited personality statuses are conferred on corporations and certain other collective entities whose legal rights are mainly limited to economic relations.

⁸⁷ A similar point has been raised by Tomasz Pietrzykowski, who recognises an intermediate category of legal entities which he calls nonpersonal subjects of law, in (Pietrzykowski 2017).

Still more limited personality statuses may be granted to other creatures, such as unborn children and nonhuman animals, and possibly also to natural entities, such as mountains, rivers, and ecosystems. On the interest-based conception of subjective rights, such creatures and entities may be granted legal rights to the extent that the legal system assumes that such creatures and entities have interests of their own that need legal protection (even though the exercise of such rights requires the activity of human agents).⁸⁸ But one can also give a *relationalist* justification to the same claim, as will be seen below.

We would not reach the same conclusions if we followed will- or choice-based conceptions of subjective rights, for which the function of rights is to ensure the holder the possibility of choosing whether or not to enforce his claim against the duty-holder.⁸⁹

In a sense, the thick personality is the one that raises more barriers between entities (e.g. between humans and nonhuman animals), while the thin one is a box that can be more easily filled.

Differences linked to the ‘thickness’ of the personality status have already been addressed in Part One (Chapter III, sec. 5.3) with regard to the forms of personalities that may be conferred on AIs, e.g. dependent or independent personalities. Indeed, the thinnest notion of personality we can think of is the one consisting only of the positions we have previously defined as *passive*, e.g. claim-rights, and typically associated with the so called ‘dependent’ personality. As MacCormick also points out: “[...] we shall note again that possession of some at least minimal legal capacity or capacities is of the essence of personateness in law. To be a person is to be able to suffer and to act in law, to act, and to be acted upon. Capacity can therefore be differentiated into passive and active capacity. Pure passive capacity is the condition of being eligible to receive the protection of the law *for one’s own sake* rather than as a means to some other end for its own sake. It is thus the minimum legal recognition that can be conferred on any being, and is that enjoyed by very young children and by incurably weak-minded persons. Other passive capacities, the transactional ones, concern rather the ability to be beneficially or onerously affected by legal acts” (MacCormick 2007, 94).

⁸⁸ The first formulation of the interests theory of rights is probably due to Rudolph Jhering, in (Jhering 1852); Some modern versions are those of Joseph Raz and Matthew Kramer: (Raz 1995) and (Kramer 2013).

⁸⁹ A well-known version of the will theory of rights is that formulated by Herbert Hart in (Hart 1984). For more recent contributions see also (Van Duffel 2012).

Yet it is not obvious that the thinnest status of personhood consists only of bundles of passive legal positions, as was the case with animals in medieval trials or slaves, who were holders of proper responsibilities but not entitled to any of the other claim-rights of legal personhood.⁹⁰

We can also separately consider the different personality statuses that certain entities may have in different areas of the law, such as company, family, medical or criminal law. For instance, a topic often debated is whether corporate entities, while having legal personality under property and tort law, also have personality under criminal law, or whether they have personality under data protection law (having data protection rights).

Lastly, differences between personality statuses may be due to the nature and purpose of the entity holding the rights and duties: the most obvious one is the different legal status of natural persons and legal persons; but another type, almost a *tertium genus*, occurs in all those cases in which legal personality has an operational rationale, often for managing goods and services, in a manner quite unrelated to the underlying individual or collective substratum, which may also be almost absent, e.g. all cases of memberless or single-member companies.⁹¹

In conclusion, the function of each personality status can be said to consist in singling out a set of rules (possibly within a specific area of the law) that apply to the corresponding type of entity, so that each instance of that type will be conferred the rights and duties deriving from such rules.

Though there are multiple personality statuses, there is a *core meaning* to which we usually refer when speaking of legal personality without further qualifications. This is the general capacity to be entitled to, given generally established triggering conditions, legal positions (rights, duties, liberties, immunities etc.). This status is shared by humans and corporations, though humans also possess further legal positions, such as human rights, the ability to enter into family relationships, and their being subject to criminal law (the extent to which such rules may apply to corporate entities is debated and is recognised in certain legal systems only to a limited extent), (Leigh 1982; Colvin 1995).

⁹⁰ Visa Kurki refers to this type of status as a 'purely onerous personality', in (Kurki 2019, 146).

⁹¹ This is not only a modern application, as we know that already in Roman law the notion of person could also refer to mere aggregates of rights and duties, for example the so-called *haereditas jacens*. See, among others, (Trahan 2008).

7. What kind of concept is legal personality?

Law is composed of a large number of different concepts. As mentioned, some of these are not purely technical-systems concepts, with a high level of abstractness – e.g. human dignity or property; others are at a lower level of abstractness and appear significantly shaped by the legal rules – e.g. contract or negligence; finally, less and less abstract concepts, perhaps included in properly technical rules, are associated with a lower level of discretionary legal determination (von der Pfordten 2009).

Within this scheme, legal personality seems to fall among the more abstract legal concepts however, taken in its proper technical sense – the capacity to be the subject of rights and duties in a legal system (Smith 1928) – it could also fall among less abstract and more legally determined ones, such as the concept of contract, offence, negligence, will, marriage etc.

In any case, the abstractness that characterises these legal concepts promotes a certain degree of openness to interpretation and favours adaptation to novel situations. This depends on a further element of many of these concepts, namely the fact that they are *cluster concepts*. In general, these are concepts – or properties – which can be applied to multiple, often disjointed, situations. In Wittgenstein's words, concepts of this genre contain: “a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail” (Wittgenstein 1953, 31).⁹²

In law, cluster concepts typically occur as sets of bundles of rights and duties that may originate in diverse incidents and unfold according to the context of reference. In this sense, legal statuses may be seen as cluster concepts as they can be obtained in disjointed ways and give rise to partially divergent consequences (Hage 2009). Not surprisingly, legal personality is often framed in the literature as a cluster concept. Richard Tur, for example, describes the personality as a cluster of legal positions from which it may result that: " [...] two entities, both of which are legal

⁹² Here Wittgenstein refers to 'games' . On the link between this conception and cluster terms see also (Parsons 1973).

persons, might have no rights or duties in common at all” (Tur 1987, 121). Similarly, Ngaire Naffine, contends personality to be a cluster concept: “[it] is made up of a cluster of things: specifically, it comprises single or multiple clusters of rights and/or duties, depending on the nature and purpose of the particular legal relation. Rights and duties, which effectively make the person, can come in thick and thin bundles, in larger and smaller clusters, which means that we are actually different legal persons in different legal contexts”(Naffine 2009, 46).

This view has recently been challenged by Visa Kurki, for whom this way of understanding cluster concepts – especially in relation to legal personality – does not conform to the standard sense: “The phrase denotes standardly a property whose extension is determined based on a weighted list of criteria, none of which alone is necessary or sufficient” (Kurki 2019, 93). Consequently, legal personality would be a cluster property in the sense that it would consist of dissociable, albeit somewhat interconnected, incidents and would be articulated in a gradual rather than the binary manner, i.e. without a sharp boundary between personality and non-personality.

Although Naffine lacks a clear reference to an analytical definition of cluster concepts, Kurki's position is no less controversial: both in identifying the standard definition of cluster concepts and applying it to legal personality (together with Naffine). There is in fact at least one other way of defining cluster concepts, which is also traditional, and that is provided by David E. Cooper: "The essential characteristic of a cluster concept, supposedly, is the following: while it is possible to list sufficient conditions for the applicability of a cluster concept term, it is not possible to list any necessary conditions for its applicability” (Cooper 1972, 496). Under this reading, cluster concepts may have sufficient conditions even when there are not necessary ones.

However, quite apart from the disputes about what is the standard meaning of cluster concepts/properties, I reject the reading of legal personality as a cluster concept subscribed by Naffine and Kurki. I don't think that from the existence of disjointed incidents triggering legal personality it can be inferred that there are no necessary and/or sufficient conditions for the application of the status. Indeed, the *entitlement* to legal positions can be seen as a necessary and sufficient – albeit quite general – condition for applying legal personality. Given that personality is (P) and entitlement is (E), entitlement is a necessary condition because if it is not true that x has the entitlement to legal

positions (Ex) then it is not true that x has personality (Ex) ($Ex \supset Px$). Similarly, entitlement is also a sufficient condition for the application of legal personhood because if Ex is true then it will be true that Px ($Px \supset Ex$).⁹³ Or also, for instance, being human could be seen as a sufficient condition of personality status.⁹⁴

Some might object that under these necessary or sufficient conditions we do not reach much conclusion on the status of a person in law. They would be right: of course, these necessary and sufficient conditions of personhood provide just a core meaning of the status of a person in law, and the configurations of the bundles of legal positions are multiple and more specific; in this sense the necessary and sufficient conditions may describe, as Naffine seems to imply, what we have also here labelled as the *thin* conception of personhood.

The concept of legal personality must therefore be seen as a classic concept, with clear-cut membership criteria – everything that does not imply entitlement to legal positions cannot be seen as legal personality – at least as regards the minimal content just illustrated. In this sense, the relationship between minimal content and the various conformations of legal personality – e.g., civil law, criminal law, or medical law personality – is a hierarchical relationship of general-to-special (*genus-species*). *fuzziness* from that of *generality*.

Unlike cluster concepts, therefore, membership of the main type is not fuzzy. A fuzzy concept is one that has no clear-cut referential boundaries: “[...] the transition between membership and non-membership is gradual rather than abrupt” (Zadeh 1965, 340). In this sense Kurki considers cluster concepts like legal personality to be fuzzy.⁹⁵ The generality of a term or concept is caused by its unspecificity – according to Kit Fine, a lack of content – which does not result from the absence of meaning or sharp membership criteria, but rather from

⁹³ It could be said that, in addition to entitlement, other contextual conditions must be met (e.g. that a centre of interest is identifiable), but in any event it is possible to say that entitlement is a necessary and sufficient condition at least for a general (or core) concept of legal personality.

⁹⁴ It could also be argued that these necessary or sufficient conditions are not part of legal personality as an institutional concept: entitlement, for example, is a meta-institutional concept of a teleological, or at most intermediate, nature, but in any case not strictly legal. This will not change the substance of my argument.

⁹⁵ In fact, the author writes: "There is no exact border between legal personhood and nonpersonhood" (Kurki 2019, 94).

the absence of details; thus, for instance, the item 'friend' is an example of generality, i.e. not falling into a specifiable category as its sex, age or nationality are not defined (Zhang 1998). Fine clearly expresses the difference between fuzziness and generality (and ambiguity) (Fine 1975). Considered two clauses:

(1) n is nice₁ if and only if (iff) $n > 15$; and n is *not* nice₁ iff $n < 13$

(2) n is nice₂ iff $n > 15$

we may claim that nice₁ is *fuzzy*, because we do not know whether it denotes the grey area in the range between the two values, whereas nice₂ is *general* since any value above that value is potentially definable by the expression.

A divergence to keep in mind is that while fuzziness is an inherent 'problem' of referential meaning of some expressions – e.g., baldness – and which is essentially unresolvable (referential opacity), generality is more a matter of interpretation or sense attribution – these senses are in any case not mutually contradictory – which is resolvable through the elimination of readings incompatible with a certain context (Zhang 1998). The relationship between the general concept and its specifications can be seen as the relationship between *genus* and *species* (or between types and tokens).

In view of the above, I consider legal personality to be a (classic) *general* property which can be specified by reference to disjointed clusters of situations; contrary to Naffine's and Kurki's thesis of personality as a cluster concept.

Moving on, we can have two different readings of meaning and reference of legal concepts: one inferentialist/reductionist and one compositional/ontological. The former reduces these concepts to rules of inference between the universe of disjointed incidents and implications; the latter sees their meaning as derived from a set of lexical components, with systems of rules – both constitutive and regulatory – as semantic referents; so that we can build a proper conceptual ontology.

I will argue that both views present advantages for understanding the function of concepts in law. However, I will ultimately claim that the compositional/ontological is the most suitable approach to describe the law as a socio-institutional phenomenon, in accordance with the metaphysical explanation.

8. The nature of legal concepts I: inferentialism

Law demands that we select facts from social practice and then reformulate them according to its own internal patterns. Thus, it creates an independent store of knowledge, so that the legal community can effectively coordinate both in implementing rules and in resolving disputes (Croce 2012). This function of the law fits legalist assumptions: legal concepts are constituted by the rules that establish their use conditions and their deontic consequences. In view of this, it is possible to conceive of legal concepts — and so also of personality — as mere links between factual preconditions and normative effects. They are basically ‘means of presentation’ of multiple rules which don’t carry any independent meaning from the link they hold, as Alf Ross believed was the case with all ‘intermediate legal terms’ (Ross 1957; Ross 1960). In short, Ross builds an example on the fictional case of a South Pacific tribe, the Noit-cif tribe, to address the issue of the meaning of such intermediate legal concepts (focusing on ownership). Any member of this tribe who engages in misconduct, such as killing a totem animal or eating the chief’s food, becomes a *tû-tû*. This status prevents the *tû-tû* from taking part in tribal ceremonies without first undergoing purification rituals. Ross’s point is that there is not really a thing in the world that corresponds to the status of *tû-tû*. Ross’s point is that the word *tû-tû* lacks a semantic referent, i.e. there is no real thing in the world that would correspond to it. Its only function and usefulness would be to hold together the set of preconditions that make someone a *tû-tû* with the effects of this status; without presupposing any further meaning. Similarly, this would be the case for intermediate legal terms, e.g. property, marriage or personhood.

This standpoint is consistent with a form of semantic inferentialism, for which, at least in Robert Brandom’s notorious version, the content of an expression coincides with (or supervenes to) its inferential role. On the side of legal concepts, this means that, rather than identifying them by their referential function, we can do so by their inferential role: by the role they play as antecedents or consequents in legal reasoning.⁹⁶ On this view, the content of a concept consists of the bundle of

⁹⁶ For an introduction to inferentialism (Sellars 1953; Brandom 1994; Brandom 2001).

inferences endorsed in its practical use and context. In Brandom's words: "[...] the content to which one is committed by using the concept or expression may be represented by the inference one implicitly endorses by such use, the inference, namely, from the circumstances of appropriate employment to the appropriate consequences of such employment" (Brandom 2001,62).

As anticipated (sec. 2), this approach to concepts is consistent with a more general normativist view about legal conceptual texture.

Hence, to employ a concept is tantamount to being committed to the inferential relations it involves with other concepts — relations such as exclusion, derivation, and implication. From an inferential perspective, legal personality links a set of triggering conditions — e.g., being human — with a bundle of legal norms — e.g., the right to sue and the liability to be sued. As a consequence, legal personality can be viewed as an implicit condition in every rule conferring rights or obligations. For instance, once the personality presupposition is made explicit in:

- If x wrongfully harms y , then x has an obligation to make y whole.

the rule should read as

- If x is a person, and y is a person, and if x wrongfully harms y , then x has an obligation to make y whole.

Inferentialism about legal concepts is consistent both with the legalistic view of personality as an abstract label that makes scattered norms homogeneous and with the idea that legal concepts embed information functional to the practical cognition of social agents. Here there are several relevant implications for the representation of legal knowledge we can benefit from.

First, identifying the meaning of legal concepts with the inferential network can be an effective way to portray the articulation of a classic concept having disjunctive preconditions and multiple legal effects, like legal personality. Second, this vision emphasises that legal concepts are meant to provide the inferences a particular legal system makes viable. Thirdly, by contextualising the legal concept in its own inferential network, one can better understand in which way and with which role the same concept is used in different legal fields. In this sense, the

inferentialist approach can also benefit interpretative activity (Canale and Tuzet 2007). Finally, by viewing conceptual links as defeasible inferences, inferentialism can enable a flexible reading of legal concepts, i.e., where the ascription and implications of such concepts are subject to exceptions (Sartor 2009).

Hence, an inferentialist approach may help us to figure out the legal effects of viewing AI systems (or other nonhuman entities) as having legal personality in the context of specific legal systems, and to find appropriate arrangements by establishing the inferential links involved. Yet, the inferentialist view on legal concepts poses some problems.

First of all, it is doubtful whether this reductionist approach can be extended to *all* legal concepts. For instance, some legal concepts – e.g., marriage – derive their meaning not only from rules, but also from semantic fields outside the law. Other legal concepts, instead, are defined more by the interdependence with other technical concepts than on rules, as von der Pfordten points out: “[...] the concept of ownership is not only shaped by the rules which attach normative consequences to it, but also by more abstract concepts like subjective right, and more concrete concepts like ownership of immovable and movable things, or ownership of corporal and non-corporal things – if this is acknowledged by a legal system. For example if one thinks that ownership of non-corporal things is possible, the relation between owner and owned thing cannot be a purely physical relation” (von der Pfordten 2009, 32).

Against this background, with regard to our ontological framework, the idea that legal personality is an inferential node is not exhaustive, as it only captures (a) its ‘use conditions’ and (b) its ‘legal implications’, but fails to include (c) the ‘background reasons’ that are part of a comprehensive theory of legal personality.

Indeed, from a merely inferentialist perspective, legal concepts do not have any intrinsic dynamism: their constitutive elements are fixed by the inferential links established by positive law. But that is not how the law actually works. A coevolutionary dynamic is at play between social and legal practice: on one hand, legal concepts are permeable to social facts — whether moral, political, or economic — and constantly adjust to their changes (Deakin 2015); on the other, the law provides us with reasons for action, incentives, and motivations, giving rise to extra-legal phenomena. Defeasible inferences that originate from disjointed events and give rise to various legal effects are held together by porous conceptual categories, not fully reducible to legal rules, which also reflect

nonlegal knowledge. A wider conceptual framework should enable legal inferences to be applied and interpreted contiguously with the social and moral landscape in which law is embedded. Integrating legal practice among the sources of inferential patterns may give a more accurate picture of the interaction between legal and extra-legal reasons, but would still not fully address the problem, which would only be shifted to another moment: the background of legal practice.

This socioempirical context places strong constraints on how rules are formed and how social agents make use of them. Figuring out how this happens is tantamount to understanding what the 'background reasons' of legal personality are and what role they play in shaping specific legal prerogatives.

An interesting critical analysis is offered by Giovanni Sartor, who, while recognising that the meaning of legal concepts can be specified on the basis of the rules that link conditions of applicability to the effects of their application, stresses how a radically inferentialist reading runs the risk of confusing the mere possession (comprehension) with the application (endorsement) of legal concepts (Sartor 2009). In order to disentangle possession and application, Sartor resorts to Frank Ramsey's thesis of the elimination of theoretical terms (in scientific theories) and Rudolf Carnap's thesis of the conditionalization of the same terms.

In short, Carnap's idea is that the analytical content of theoretical terms can be expressed through conditional sentences: i.e., *if* there is a category, or a relation, e.g. patriarchal marriage, that satisfies the inferential links embedded in a concept, *then* we can assume that this category is the same as that denoted by the term defining the concept.⁹⁷ In this sense, accepting Carnap's conditional statement involves possessing the concept in question, e.g. patriarchal marriage.

Ramsey's idea is instead that theoretical terms, i.e. those that recur in scientific theories (e.g. energy), can be eliminated and replaced by

⁹⁷ Sartor exemplifies this conditional statement as follows: "*if there exists a category Z such that*

– IF a couple goes through a marriage ceremony, THEN they are in this relation Z, and

– IF a couple is in relation Z, THEN the husband has power over the wife

then

– IF a couple goes through a marriage ceremony, THEN they are in relation patriarchal marriage, and

– IF a couple is in relation patriarchal marriage, THEN the husband has power over the wife", (Sartor 2009, 232).

existentially quantified variables. Similarly to what scientific terms do, connecting empirically observable data to predictions of phenomena, legal concepts connect observable facts to normative consequences (and qualifications). The value of Ramsey substitution is that it: “[...] makes the assumption explicit that there exists some predicate that, substituted for the variable, yields true or valid propositions” (Sartor 2009, 230). And this implies a different kind of relationship with the concept in question since: “Having a concept would not only consist in understanding an idea but would also amount to claiming that this idea has concrete reality, namely, that there exists a category for which the concept’s constitutive inferences really obtain. For instance, possessing a concept of patriarchal marriage would entail assuming that there exists (in the context of the legal system we are considering) a relation Z having the following inferential features: (a) if a couple goes through a marriage ceremony, they are in this relation Z, and (b) if a couple is in the relation Z, then the husband has power over the wife. The existence of the category entails the holding of its inferential links” (Sartor 2009, 231). In other words, Ramsey’s elimination implies not only possessing or understanding the theoretical term but also endorsing its actual application; or as Carnap argued, expressing the factual content (Psillos 2000).

The application of legal concepts, therefore, differs from their mere possession/understanding in that it additionally implies a kind of doctrinal commitment, i.e. the belief that the inferences constituting the concept actually hold in a given legal system. The reasons why these inferences do or do not hold in a legal system suggest that a deeper analysis of the meaning of legal concepts should be conducted, perhaps calling into question a true conceptual ontology – which also has its limitations – accounting for a better understanding of law as a social and moral practice.

This is not to say that the inferentialist approach is nonsensical or invalid. As Sartor points out: “[...] an inferential theory of legal concepts cannot limit itself to a strong (committing) way of having a concept; it must also provide a weaker way of possessing conceptual inferences, a way not involving the acceptance of such inferences, namely, not involving the belief that the corresponding inferential links hold in the domain we are considering. Such a weaker kind of inferential understanding should enable us to knowingly possess defective concepts (concepts whose constitutive inferential links we know to be wrong), or

local concepts (concepts whose inferences we know to be applicable in certain contexts, but not in certain other)” Sartor 2009, 228).

In any case, I think that one of the main limits of the inferentialist view is that it leads to thinking that the reduction of complex bundles of legal rules (and relations between them) to unitary packages held together by single terms is somehow arbitrary. Put another way, inferentialism would not account for the ‘background reasons’ of legal personality, i.e., it would not explain why we group different situations under this doctrinal category and how extra-legal information is integrated into it. I do not believe that the gathering of incidents and rules within a single concept-term is accidental, but rather that it is anchored in the origin of institutional (legal) practice and the promotion of certain goals and conduct; these elements may come from domains outside the law.

To conclude, I maintain that it would be preferable to subscribe to a less demanding version of inferentialism: even though legal concepts are (partly) constituted by their inferential role — by the set of rules authorizing their inferential use — this would not be sufficient to account for their axiological or teleological fitness.

9. The nature of legal concepts II: ontology and neo-institutionalism

In this section, I counter two of the cornerstones of inferentialism, namely (a) semantic reductionism and (b) the lack of semantic references. I believe that these two theses can be contrasted in an organic way, with counter-theses grouped into a single theory of concepts of a systematic-ontological nature. Indeed, this view associates conceptual categories in various ways with terminological definitions – and relationships between terms – and accordingly assign meaning to them, giving rise to conceptual architectures also known as *ontologies*.

A neo-institutionalist model of legal concepts will be preferred, mirroring a certain way of understanding legal reality (as emphasised in the previous chapter), and from which we will frame the concept of legal personality. This neo-institutional ontology, here presented in its more or less traditional version, will be later updated to include further layers of explanation of legal concepts, and hence of personhood.

9.1. Compositionality

An alternative approach to inferentialism regarding conceptual meaning is that expressed by the so-called compositionality thesis.

In general, according to proponents of compositionality, the meaning of expressions, rather than being constituted by the set of propositions that can be inferred (as true) from them, is constituted by their lexical components and the relationships between those components. A comprehensive definition of the principle of compositionality, also called Frege's principle as the German philosopher is one of the first exponents of this thesis, may be: “[...] the meaning of an expression is a function of, and only of, the meanings of its parts together with the method by which those parts are combined” (Pelletier 1994, 13).

As compositionality tends not to be tied to specific theories of meaning, it mainly constrains the meaning of complex expressions, i.e. a collection of semantic parts, and requires that the meaning of the collection be derivable from the meaning of its parts. Thus, contrary to the inferentialist approach, the meaning of the parts (terms) is not derived from the whole (sentences and inferential links), but the other way around.

Our interest is not in language or meaning in general, but, more specifically, in concepts, especially legal concepts. So, what does it imply to take a compositionalist view on concepts? As Sartor makes explicit: “Conceptual knowledge is packed into the terminology, and is expressed through the definition of terms, and through the specification of connections between terms. Rather than abstracting terminological meaning from sentential inferences, we express a conceptual framework through a terminology, and then use this conceptual framework to express substantive information” (Sartor 2009, 236). However, this still does not provide sufficient arguments as to whether compositionality actually captures the nature of concepts and/or their functioning.

One of the most accurate insights into the compositionality of concepts has been offered by Jerry Fodor (Fodor 1998; Fodor 2002). Fodor adheres to those theories that identify concepts with mental representations – representational theories of the mind (RTM) – whereby a concept is any mental particular that “[...] satisfy whatever ontological conditions have to be met by things that function as mental causes and effects” (Fodor 1998, 23).

According to Fodor, mental representations, and so concepts, normally have the characteristics of compositionality, in the sense that their content is determined by the content and syntactic structure of their primitive constituents.⁹⁸ To argue this position, Fodor resorts to some traditional justifications: compositionality must be true since it is the only thesis that can account for the *productivity* and *systematicity* of human linguistic and cognitive aptitudes.

In particular, the productivity argument postulates that although the representational capacities of human beings are finite, the number of concepts that an individual can grasp are almost infinite.⁹⁹ In fact, a competent speaker potentially has the ability to understand a new expression 'T' without further information about its meaning. If this is the case then the speaker has prior knowledge, namely the structure of T and the meaning of its individual primitive constituents.

The other argument in favour of compositionality of concepts (and thoughts), namely that of systematicity, postulates that the sentences we understand occur in predefined and predictable patterns: "The reason that a capacity for *John loves Mary* thoughts implies a capacity for *Mary loves John* thoughts is that the two kinds of thoughts have the same constituents; correspondingly, the reason that a capacity for *John loves Mary* thoughts does not imply a capacity for *Peter loves Mary* thoughts is that they *don't* have the same constituents. Who could really doubt that this is so? Systematicity seems to be one of the (very few) organizational properties of minds that our cognitive science actually makes some sense of" (Fodor 1998, 98).

Fodor also provides an empirical argument in support of the systematicity of thinking and has to do with how children learn expressions and concepts - who typically learn the meaning of new concepts from context or from interlocking with other expressions - but unfortunately we cannot dwell on here (Fodor 1998).

To sum up, Fodor's argument is that the compositionality hypothesis is true if and only if the constituents of concepts determine their syntax and content in a way that explains their productivity and systematicity.

In any case, what is relevant to our interest is the interconnection between compositionality and the ontological approach to concepts,

⁹⁸ In particular, Fodor rejects the so-called theory of concepts as prototypes (Fodor and Lepore 1996).

⁹⁹ An early formulation of the productivity argument can be found in Frege.

especially legal concepts. Ontology in a sense presupposes compositionality. If it is true that concepts are compositional, then it is possible to construct conceptual ontologies as a function of the various types of relations in which they can be found: e.g. relations between primitive and complex concepts, between constituents and constituted, relations of species and genus (types and tokens), mereological relations, relations of priority or necessity, and so on.

It seems to me that a case of compositional content analysis is offered by Himma in reference to the concept of 'bachelor', i.e., as composed of further concepts such as 'man' 'adult' and 'unmarried' (Himma 2015). What does it imply, instead, to regard legal personality as a compositional concept? It implies, to a first approximation, that its meaning may result from lexical components such as those of, say, legal status, legal position, entitlement, claim, obligation and so on; or even more abstract ones such as matter, spirit, body, reason, sentience and so on. I will identify an ontological distinction between these two classes of concepts.

9.2. Institutional reality and legal personality

Another main tenet of inferentialism about concepts posits the absence of any semantic referent, which means that there would be no phenomenon of the space-temporal world corresponding to intermediate legal terms. As Ross argues in 'Tû-tû', these terms would have no reality and would only serve to hold together chains of legal rules and arguments. Yet this argument too can be contested.

In fact, as we saw in the previous chapter, there is at least one way of seeing legal concepts as having semantic referents, namely those entities that are part of social and institutional reality (the latter is a subset of the former). The reference to socio-institutional reality to substantiate the existence of semantic referents of legal concepts has also been highlighted by Jaap Hage, with a special focus on the so-called legal status words (Hage 2009).

Of course, these are not physical entities, but are what have previously been designated as socio-institutional kinds and facts; such entities, especially in the form of institutional facts, certainly have a temporal (and spatial) dimension.

Legal concepts – at least the nodal concepts – can appear as subspecies of these socio-institutional kinds, and it is possible to associate them with a certain metaphysical framework. Depending on

which theory of social ontology one intends to subscribe to, and thus on what is meant by institution, one will refer to a somewhat different content. As I shall argue in the next chapter, the view of institutions as a set of rules in equilibrium proposed by Hindriks and Guala is quite convincing also for legal facts and kinds. Therefore, the reality and the referents to which the legal concepts refer are precisely the facts that are determined by these rules (in equilibrium): “A fact exists as an institutional fact if (and only if) there exists a rule in social reality that attaches the existence of this fact to the existence of other facts [...] For instance the fact that John is punishable is an institutional fact. It is attached to the fact that John is a thief by the rule that makes thieves punishable. Another example would be the existence of money and the possibility to use money for making payments. The existence of money and its use is presently regulated by the law, which is a phenomenon that exists in social reality itself” (Hage 2009, 60). Trivial though it may seem, these rules increase the number of facts that recur in the world.

In my view, this approach enhances Searle's institutional theory while broadening the spectrum of rules and functions performed by institutions, at the same time making it more compatible with MacCormick's legal neo-institutionalism. The latter is one of the best-known readings of legal concepts as (rule-based) institutions – labelled as institutional legal concepts – and for this purpose, I will take it as an alternative model to a rigidly inferential approach. Yet, even if this conception of legal institutions is rule-based, the way MacCormick characterises their origin and the type of rules that constitute them shows a certain gap with Searle's canonical reading.¹⁰⁰

MacCormick, by the way, provides a critique directed against inferentialist reductionism of legal concepts. He discusses the classic objection that intermediate legal concepts, which he calls institutional, can be eliminated and replaced by a single rule that holds together use conditions and legal consequences, which for personality might sound like “if two or more persons associate for a common purpose, and share part of their assets for that purpose, then they acquire certain rights and duties jointly”.

¹⁰⁰ MacCormick will eventually abandon this reading: he preferred to conceive of legal institutions as autonomous normative systems, somewhat following the institutionalist tradition à la Santi Romano (MacCormick 2007).

This objection, although theoretically well-founded, has, according to MacCormick, practical limitations. In fact, jurists and legal officials when using conceptual categories tend to separate the use conditions from the effects caused by the existence of the single institutional instance. The need to separate complex normative sets, e.g. the law of contracts or the law of persons, into simpler units of rules justifies this attitude towards legal concepts: “ [...] the use of the use of the concept of ‘contract’ enable us to achieve the desirable goal of rendering the rule into two simpler unitary rules. [...] This leads on to a vitally important observation from a jurisprudential point of view: it makes it possible to state as separate legal rules a legal provision which confers legal power and one which imposes a legal duty. By using the notion of the ‘existence of a valid contract’ we can conceive of two separate rules, the one which enables a class of people (those who have ‘contractual capacity’) by certain acts to bring a contract into existence, and the one which ordains that those who have done so acquire primary rights and duties” (MacCormick 1986, 59).

In my view, this argument, besides having pragmatic-ontological value, can also underpin the thesis of compositionality insofar as it sees the simpler regulatory units as semantic components of concepts, rather than as inferential chains from which the meaning of the concept is deduced.

In the next section, I take a closer look at the neo-institutionalist conceptual framework, looking at what rules make up institutional legal concepts and how this architecture may be applied to legal personality.

9.3. A Neo-institutional account of legal personality

Against this background, to move beyond inferentialism strictly understood and also to integrate the thesis of compositionality with that of institutional reference, we will refer to the theory of institutional legal concepts developed by Neil MacCormick.

Drawing on Searle’s social ontology, MacCormick views legal facts as institutional facts (MacCormick 1986). An institutional fact is conceived as something whose existence depends ‘not merely upon the occurrence of acts or events in the world’ — as is the case with brute facts — but

also upon rules.¹⁰¹ Indeed, institutions, from which institutional facts descend, presuppose a system of rules of a peculiar kind: constitutive rules. In a nutshell, constitutive rules generally have the logical form ‘X counts as Y in context C’, where Y stands for the *status function* these rules assign to people or objects within a given context C (Searle 2009). The assignment of status functions, Searle claims, cannot be explained by the physical makeup of those people or objects but is rather the result of collective intentionality, such as our disposition to cooperate and share desires and beliefs about the institutional condition of certain entities. The status so recognized triggers further consequences, which Searle calls ‘deontic powers’: e.g., rights, obligations, requirements, permissions, authorizations, entitlements. Money, for instance, is an institution insofar as it presupposes a system of constitutive rules assigning an agreed exchange value to certain objects (X will count as money in C); defined functions will correspond to those objects and a series of powers and duties will be entrusted to those who use them.

Therefore, MacCormick maintains that systemic legal concepts — like property, marriage, succession, and, of course, personality — are institutions of law whose specific instances are institutional facts. For instance, a company having a legal personality is an institutional fact that relies on the institution of legal personality. Such concepts, which he calls *institutional legal concepts*, are systemic in that they hold together many interrelated legal rules. They are used, in fact, as tools for sorting out intricate normative content.

From MacCormick’s perspective, a legal institution can then be approached by bringing three distinct things into focus. Firstly, we may focus on the institution itself, i.e., institutional legal concepts, as the set of (constitutive and regulative) rules. Secondly, we may focus on single instances of the institution, i.e., institutional legal facts, that are created by concrete behaviours and events appropriately matching constitutive requirements. Thirdly, we may focus on the social practice supporting the institution, namely, on the social and individual interests the institution is meant to satisfy, and on the individual and collective action aimed at such interests (MacCormick 1988). At this point in time, as repeatedly stated, it is the institution as a concept that interests us — although we are also concerned with how the theory organically holds

¹⁰¹ This distinction between brute and institutional facts can originally be traced back to Gertrude (Anscombe 1958).

these levels together – and which MacCormick often emphasises (also through the notion of *arrangements*): “[...] it is better to use the idea of an "institution" to signify the conceptual framework within which particular arrangements can be set up by particular persons on particular occasions to last through particular periods of time” (MacCormick 1988, 76).

As previously mentioned, according to MacCormick, each institutional legal concept is regulated by a triad of rules:

(1) *Institutive rules*. Rules linking the occurrence of certain acts or events to the coming into existence of a specific instance of a legal institution: e.g., the rule stating that if a group of people allocate money to a common asset for an associative purpose, then an association having legal personality comes into existence (Ruiter 1993).

(2) *Consequential rules*. Rules determining the implications of the existence of each specific instance of an institution: e.g., if a legal person exists, then this person can be held responsible for any loss, damage, or injury it causes.

(3) *Terminative rules*. Rules laying down the conditions under which a specific instance of an institution of law ceases to exist and thus ceases to produce legal effects: e.g., the legal personality of a collective body dissolves when the purpose of the association becomes unattainable.

Unlike Searle, MacCormick does not regard this triad of rules as constitutive rules, but rather as a set of constitutive and regulatory rules: “The 'constitutive rule' as Searle envisages it includes part of both 'institutive' and 'consequential' elements as I account for these. And it fails to convince as a 'rule' in the normative sense, that is, as a guide to judgment and conduct. For the boundary between regulative and constitutive is unclear in Searle's schema. In fact, the full panoply of institutive, consequential, and terminative rules in my schema is 'regulative' in the sense of regulating (or at any rate authoritatively guiding) how people are to undertake commitments when they think fit to do so, in what spirit they must honour their commitments made, and how they are in the end to wind them up or be released from them” (MacCormick 1998, 335).

However, Searle's and MacCormick's positions are less conflicting insofar as we accept the existence of two types of constitutive rules: *by condition* and *by implication*.¹⁰² Basically, this means that institutional concepts — like money and borders — but also legal ones — like personality, contract, and ownership — are constituted not only by their conditions of existence but also by their consequences. In other words, the constitution of an institutional practice could not be said to be completed without setting out the effects and implications, perhaps in the form of deontic powers, that it produces with respect to the subjects and the reality that hosts it. In this way, MacCormick's institutive and terminative rules would appear as constitutive *by condition*, while his consequential ones as constitutive *by implication*.

Hence, legal personality can be conceived as an institutional legal concept: a systemic concept made up of institutive, consequential, and terminative rules. This way of conceiving legal personality yields an account of the temporal persistence of its specific instances, which come into existence as a result of certain acts or events, operate through sets of legal implications like rights, duties, powers, and responsibilities ('status functions' and 'deontic powers') (Searle 1996), and finally end at a given moment.

MacCormick's view is somehow compatible with an inferential reading of legal concepts since the institutionalist architecture looks consistent with the legalist idea that personality is a label through which the legal system organises fragmentary rules in order to channel them toward a single point of imputation. This compatibility seems possible at least with a *weak* version of inferentialism.

¹⁰² This idea was suggested by Corrado Roversi under the name of 'complementarity thesis' in (Roversi 2012). See also (Hindriks 2005).

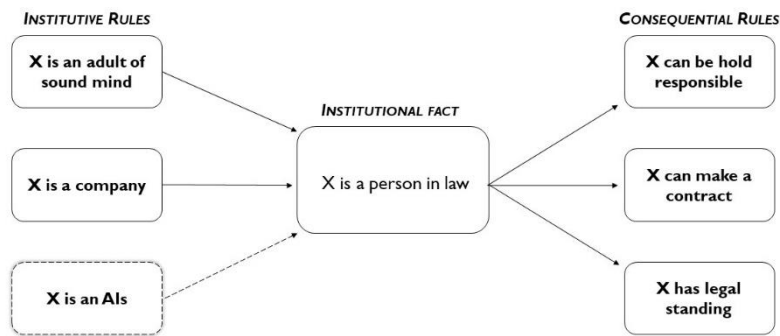


Figure 3. Inferentialism and MacCormick's neo-institutionalism on legal personality

MacCormick himself recognizes the inherent limitations of his triad of norms and in general, despite the strong influence of Searle's social ontology, of constitutive rules in providing an adequate account of institutions. After all, this skepticism appears consistent with his theoretical mission, which is to reconcile the legalistic view and the social dimension of legal phenomena. In fact, MacCormick claims that a profitable understanding of institutions would make it necessary to lay bare their 'underlying principle' or 'final cause': "[...] there is value in trying to answer the question that the so-called 'constitutive rule' answers. Effectively, as I have suggested, the question is about the point or general aim, or end, that is, the 'final cause', of any particular practice or institution. The mere fact of being storable in terms of a triad of institutive, consequential, and terminative rules is after all something that all institutions have in common. It is in their final causes, reflected, of course, in the content of the triadic rules, that they differ. Thus for each it is possible to formulate some guiding principle or principles that express the underlying final cause" (MacCormick 1998, 335).

Hence, an institution and its specific instances can only be understood in light of the general purpose they serve, in the context within which they operate. As MacCormick exemplifies, "corporations are associations of individuals to which a separate legal personality attaches for the purpose of holding property and bearing and

discharging legal obligations and responsibilities” (MacCormick 1998, 336). The point, or final cause, of institutions then also has specific value for the application and interpretation of institutional concepts (i.e. of their instances): “They are also of great importance in interpreting and applying or implementing particular instances of given institutions. Without some broad conception of what contracts or trusts are for, it is impossible to achieve an intelligent interpretation of contract law or the law of trusts, or therefore, of any particular contract-regime or trust-regime” (MacCormick 1998, 336).

Yet, MacCormick never specifies what role these underlying principles play in his (neo)institutional ontology of legal concepts. Intuitively, they seem to be traceable to the sociological dimension of institutions, since they do not seem to be subject to the same conditions and rules as the purely institutional dimension, but this reading might be limited or at least ambiguous.

At this scope, in the next section we shall further develop MacCormick’s perspective by exploring the possibility of integrating these aspects, that go beyond the usual account of institutional concepts and facts, to which end we will be relying on the path charted by social ontology. This analysis, which will be explored primarily through a clarification of the (neo)institutional conceptual framework, will also help to clarify the relationships that exist between the sociological and the purely institutional dimensions; able finally to display more accurately the architecture of concepts in law and, above all, of legal personhood.

10. Developing the neo-institutional ontology: meta-institutional concepts

MacCormick’s theory proposes different degrees of existence of legal institutions, in such a way as to also take their social dimension into account, but his theory does not single out the conceptual links between social and legal kinds. The main theoretical tool we get from the theory is that of institutional legal concepts, which, as we have seen, are suited to an inferentialist approach that does not provide room for the background of constitutive (and regulative) rules. For this reason some ideas coming from the literature on social ontology could help us update the neo-institutional ontology.

One of the most cited arguments in favour of a broad understanding of social ontology — not limited to constitutive rules — comes from the thought experiment proposed by Hubert Schwyzer, who highlighted that there is something behind the constitutive rules of an institutional practice (Schwyzer 1969). Schwyzer imagines a society —the ‘Ruritarians’— in which the game of chess exists with its typical rules, but instead of being a competitive game it is a religious ritual. As such, at the end of the game the audience is all very relieved if white chess wins over black chess because this is a sign of prosperity for the community; no one congratulates the player who "won" nor does anyone comfort the one who "lost". Also, there is only one chessboard per community and no one is allowed to play for fun.

It follows that Ruritanian chess lacks the concept of ‘victory’ or ‘defeat’ and that the actions associated with such concepts cannot be carried out. Yet the concepts of victory and defeat are logically independent of the constitutive rules of chess, since they are rooted in the wider social practice in which the game of chess is immersed, i.e., the practice of competitive game playing. These concepts would lie in the background of constitutive rules, a further level composed of concepts that are *presupposed* by institutional practices. That is, there are factors which are not specified by the rules of social practice, but which bring the institutional practices into being and, at the same time, enable the practice to function in its distinctive way. These factors are covered by *meta-institutional concepts*; The expression ‘meta-institutional (concepts)’ was first introduced by Dolores Miller (Miller 1981, 188).

Back to our topic, let us see what further levels can contribute to more adequately framing and explaining legal personality, thereby ultimately giving us a better handle on the issue of AIs. According to the inferentialist account it is possible to explicate the concept of legal personality for natural persons in this way:

(1) If x is a human being, then x has legal personality (*constitutive by condition*).

(2) If x has legal personality, then x can enter into contracts (*constitutive by implication*).

Rules (1) and (2), when applied to a specific fact (e.g., Jane Doe is a human being) converge into the same institutional fact: ‘Jane Doe has

legal personality,' which entails that 'Jane Doe can enter into contracts'. In MacCormick's terms, this fact – 'Jane Doe has legal personality' – counts as a specific instance of an institution of law. But this institutional fact also seems to depend on presuppositions that are not defined through the set of constitutive rules of legal personality. In the case here proposed, (2) seems to assume the capacity of the entities at stake (those being granted legal personality) to act in accordance with their rights and duties, to intelligibly communicate, and to rationally deliberate. The family of presuppositions of legal personality is not a fixed set but rather contains social justifications and moral perspectives that may change over time. For example, where legal personality manifests as the possibility of acquiring rights and obligations, or to sue and be sued, agency and moral competence are taken for granted. In other cases, where, for example, legal personality manifests merely as instrumental to the protection of interests (i.e., *passive* legal personality) (MacCormick 1988, 383), or values — e.g., vulnerability or integrity — are presupposed relative to the legitimacy of the protection provided by the status. These presuppositions appear to be *prior* to that of legal personality, in the sense that they condition the possibility of the institution itself (Lorini, *Meta-Institutional Concepts: A New Category for Social Ontology* 2014). For instance, the concept of agency does not originate with the practice of law itself but rather emerges from a wider context which law is part of. To some extent, it is precisely the role of law in a broader social practice that makes it so that agency should be presupposed (Roversi 2014).

This aspect seems to me to be compatible with what Naffine says about the limits of a strictly legalistic view of legal personality, specifically about the meanings that the concept of personality inherits – and presupposes – following a process of historical stratification: “But if metaphysical meanings of the person, be they Rationalist, Religionist, Naturalist, have already entered the legal lexicon, as I will suggest they have, then such metaphysical uses *are* legal uses. If, for example, judges invoke human sanctity (as they commonly do) when they are considering whether to endow any given human being or entity with rights, then Religious meanings of the person have already found their way into law. The Religious person is already conditioning the meaning of the Legalist's person and it is poor scholarship, especially according to the lights of Hart, not to consider this Religious manifestation of the person within law. For the Religious meaning is not just tempering Legal

meaning; it *is* legal meaning; or, rather, it is one of a rich variety of legal meanings. To press the point home: if we neglect this spiritual being (or Naturalist or Rationalist being for that matter), we are not attending to legal use. We are neglecting significant legal manifestations of the person” (Naffine 2009, 41).

Meta-institutional concepts linked to legal personality are not constituted by the normative conditions for being a ‘person’ in law, but rather amount to moral, social, economic or political factors. Yet it would not be possible to design the constitutive rules of legal personality or to critically reflect upon such rules without implicitly relying on some of these notions. Depending on the concept (or family of concepts) presupposed, differentiated personality statuses may be obtained.¹⁰³ The meta-institutional level somehow reveals that the law must contain, at least implicitly, some reference to the attributes of persons that place them within the scope of certain legal provisions.

So, what does the meta-institutional level consist of? It cannot be made up of institutional-type rules having the form ‘*x* counts as *y* in *C*’, since in this case meta-institutional concepts would not be distinct from institutional concepts. It also cannot be made up of mere brute facts, since in this case, it would be incapable of providing normative justifications/explanations for institutions. Rather, it is made up of social values, conventions, standards of conduct, and shared beliefs which are (at least partly) exogenous to the institutional rules; in some cases, it is the combination or historical layering of these elements that make up the meta-institutional tier. It includes different kinds of concepts: some of them are *axiological* – e.g., ‘competition’ for chess, and ‘justice’ in the legal domain, or ‘dignity’ for legal personhood – while others are mainly *teleological* – e.g., ‘victory’ in chess, and ‘validity’ in the legal domain, or ‘capacity’ for legal personhood (Roversi 2014, 205).

Not surprisingly, the concepts that make up the meta-institutional level will also change according to the theory of legal personality that one subscribes to: a spiritualist will consider the institutional concept of personality committed to meta-concepts such as the sacredness of human life; a relationalist will consider it committed to the type of connections that a certain entity entertains within society and to the type of social needs it satisfies; a legalist to the systemic and instrumental

¹⁰³ Relevant to the pluralism of determinants of personhood —both legal and moral— is Gellers’s multi-spectral approach in (Gellers 2021, 151).

reasons for the organisation of legal relations and positions; and so on. These reasons, or instrumental constraints, determine a limited application of legal personality and, more generally, of legal concepts. But these limitations can be much looser in the case of a predominantly legalistic and pragmatic reading of legal personhood, fostering the conviction that any entity can abstractly be a person of law – even an armchair – as long as the ascription of this status serves a (relevant) function in the eyes of the legal system (Banaś 2021).

Meta-institutional concepts transcend the boundaries of single institutions — hence the boundaries of the structure set up with constitutive rules — and give meaning to the institutional practice within its socioempirical environment. In the case of personality, such concepts — ideas of humanity, agency, sentience, integrity, environmental value, economic expediency, etc.— reflect individual and collective attitudes in different fields, values that both drive legal practice and provide a common ground between the legal, moral, and sociopolitical domains. The institution of legal personality can truly be ‘played out’ only if contextualized within this background that meta-institutional concepts display.

11. Conclusions

In this chapter, I have provided a conceptual analysis of legal personality, but offer arguments so that it can also serve other comparable concepts. In this sense, the theoretical exercise is what is generally referred to as, as already mentioned, conceptual jurisprudence, and it is aimed at unpacking the content of terms in order to place them in conceptual frameworks or architectures that show the kind of practices – linguistic and otherwise – in which we are involved.

In particular, I have argued that the concept of legal personality is a classic (and general) concept, rather than a cluster (and fuzzy) one, as it is sometimes claimed (sect. 3).

I then presented two readings about the essence of concepts, applying them to legal concepts: inferentialism and compositionism (Sections 4 and 5). Here I have tried to show the advantages (and disadvantages) of both readings, embracing the thesis that a weak inferentialist approach can be privileged while being compatible with the ontological-compositional approach.

I used the neo-institutionalist theory of legal concepts as a model that would hold together some aspects of compositionism and inferentialism - as well as, of course, institutionalism (also for the purpose of semantic reference of concepts) – trying to develop a neo-institutionalist account of legal personality (sect. 5.3).

Despite the merits of this perspective, however, some of its inherent flaws have also been denounced. A proposal has been made to update the neo-institutionalist theory through the use of the – both ontological and conceptual – so-called meta-institutional layer (sect. 6). In a conceptual sense, it was argued that legal personality is an institutional concept constituted by a set of (constitutive) rules, but it is at the same time constituted by practices and beliefs, collected in meta-institutional concepts, which are part of the social context from which law originates. The presence of an intermediate conceptual category between these two has been opened up, although it will be explored in more detail in the next chapter.

In the next chapter, I will show that the relationships between layers of this conceptual ontology of legal personality can be described in terms of metaphysical relationships: grounding for the institutional level, i.e., relating the kinds/facts described by legal rules to the legal kinds/facts; and anchoring, or structuring grounding, for the meta-institutional level, i.e. relating the meta-legal kinds/facts to the legal rules.

III. The Nature of legal personality: real definition analysis

1. Personality as an institutional kind: social ontology and neo-institutionalism

In this section, I will offer an institutional account of legal personality, which can also be extended to other legal concepts.¹⁰⁴ I will deal later with providing a metaphysical framework for this category. For both these purposes, I will employ an approach of social ontology, somehow a narrower field of metaphysics, to identify the foundations and building relationships of that peculiar class of social kinds of which legal personality is a part, i.e. legal kinds (Epstein 2016).

A major project of social ontology that merits incisive legal-philosophical reflection is the one pioneered by Neil MacCormick and Ota Weinberger, considered to be the founders of the neo-institutionalist current (MacCormick and Weinberger 1986). Neo-institutionalism is strongly inspired by the philosophical contribution of Elizabeth Anscombe and John Searle, who provide some of the most

¹⁰⁴ It seems reasonable to think that similar considerations to those that will be made for personality could apply to equally nodal concepts such as property, marriage, wills and so on.

influential and organic theories of the social world. From Ascombe and Searle, it will be seen, the two neo-institutionalists derive the distinction between brute and institutional facts – including the legal phenomenon in this latter category of facts – as well as the notion of constitutive rules (though not entirely agreed).

Following the neo-institutionalist theory, especially in the version advanced by MacCormick, I will describe legal personality as an institution of law, or an institutional fact, while updating this account in the light of the contemporary debate on metaphysics and social ontology. Specifically, I will recall the debate on the nature of social institutions, generally contested between a rule-based and an equilibrium-based account, and I will report on a theory that tries to combine these two approaches, that of Frank Hindriks and Francesco Guala, trying to conclude whether or not it is useful also for conceiving legal institutions.

Later, I will recall some of the arguments defended by Brian Epstein in ‘The Ant Trap: Rebuilding the Foundations of the Social Sciences’ (2015) about the nature of the social world, also drawing on some of the diagnostic tools of analytic metaphysics (e.g. grounding) with the aim of integrating a more comprehensive metaphysical account of the origins and building blocks of legal-institutional kinds, and therefore of personality, rather than settling for that provided with institutional theories of social reality alone.¹⁰⁵

1.1. Social ontology

Social ontology is the branch of metaphysics that gathers various theories about essential properties and sources of social phenomena. These investigations seek to answer the more general questions of what social reality is and how it is set up. The first, by identifying the conditions under which something counts as a social entity, e.g. what counts as a ‘State’; the second by identifying what brings it into being, e.g. whether, and how, it is determined by natural entities. Theories of social ontology thus identify a set of facts and objects whose existence is not independent of human activity – unlike natural entities – and seek

¹⁰⁵ More emphasis will be placed on neo-institutionalism with regard to the conceptual analysis of personality, and in that too an attempt will be made to update the architecture of legal concepts proposed by MacCormick.

to justify them through other facts and objects by virtue of which they occur (Epstein 2018).

There are numerous entities analysed by social ontology, and, depending on the particular approach or theory, some of them are taken as more emblematic than others to describe the characteristics of the social world. Preference may be given to artefacts, institutions, group intentionality or complex phenomena such as language and law. In doing so, social ontology intersects different disciplines: from cognitive science to action theory, from sociology to economics, from political philosophy to legal theory and more.

According to Francesco Guala, some theories of social ontology can be collected within a Standard Model due to the sharing of certain theoretical assumptions (Guala, *The Philosophy of Social Science: Metaphysical and Empirical* 2007). The premise, or 'basic ontological intuition', of the Standard Model commits, albeit weakly, to a individualistic view of the social world: "[...] if all individuals were to disappear, all social institutions would disappear too".¹⁰⁶

In addition to this individualist foundation, the Standard Model has other relevant theses. Guala stresses three of them: (1) the social world is constituted by virtue of our attitudes and beliefs, sometimes reflexive, i.e. beliefs about beliefs; (2) it has to be (re)created constantly by members of the community through linguistic performatives such as declaratives (Searle) or by the preservation of practices and beliefs (Hume); (3) it is the product of collective intentionality, i.e. shared attitudes and we-mode beliefs, like collective acceptance and conventions (Guala, *The Philosophy of Social Science: Metaphysical and Empirical* 2007, 961).

One of the theories that is founded on the claims of the Standard Model, and is unquestionably the most influential of the last twenty years, is that of John Searle. Searle's account emphasises the role of language in social ontology and uses the category of so-called *institutional facts* to describe a vast set of social facts and objects, e.g. money, borders, public offices, etc. Searle develops this notion in the wake of Elizabeth Anscombe's reflections on the existence of certain facts that are not reducible to others, i.e. brute facts, as opposed to another class of facts

¹⁰⁶ This observation is linked to the broader debate between individualistic and holistic theories of social reality (a distinction that may apply to ontological positions as much as to methodological ones); an overview of the debate is also offered in the first three chapters of (Epstein 2015).

that only recur in the context of human institutions (that make up the *institutional reality*) (Anscombe 1958; Searle 1969). In Searle's theory these brute facts correspond to natural/physical facts, e.g. fact that the Earth is 93 million miles from the sun; and not all social facts are institutional, but only those whose existence depends on systems of constitutive rules (i.e. institutions), e.g. the fact that Joe Biden is the President of the United States of America (Searle 2010, 10).

Searle recognizes three primitive factors underlying institutional reality: (1) constitutive rules, (2) collective intentionality, and (3) assignment of (status) function. Since the constitutive rules are not laws of nature, for Searle something becomes institutional by virtue of the use of specific linguistic acts, the declaratives, and a peculiar mental attitude towards those rules, i.e. the collective acceptance by the members of a certain community (Searle 1996). It is by virtue of these rules – in the logical form 'X counts as Y in C' – that entities are assigned functions and powers that they would lack by virtue of their purely physical and natural composition.

In a nutshell, social facts that make up institutional reality are created by communities through the language and collective acceptance of constitutive rules, through which objects, people, entities and events are assigned status functions that convey (one or more) deontic powers. So, a piece of paper is assigned certain functions – e.g. to be a store of value – because in the European legal system (C) it is collectively accepted the rule that every piece of paper validly issued by national central banks (X) counts as a 'euro' banknote (Y).

Searle's approach is actually akin to an older theory of social reality, also attributable to the Standard model, which is that of David Hume (Hume 1740). For Hume, the conditions for something to count as social are brought about by a combination of widespread beliefs, regularities of behaviour, and material practices (implicit or explicit) not involving necessarily a specific mental attitude such as collective acceptance; for example, the linguistic practice whereby certain propositions expressed in the form of a promise imply an obligation. It follows that Hume's theory has a more conventional root.

Also worthy of mention is an alternative view with respect to Searle's theory of constitutive rules, and one that has attracted attention in recent years, is that of Maurizio Ferraris' documentality (Ferraris 2012). To put it as simply as possible, Ferraris' thesis is that social reality can be accounted for by the rule that every social object follows from the

registration of acts involving at least two persons, and which are recorded on any physical medium: e.g. paper, marble, the web or neurons.

It is not possible to go deeper into the social ontology debate in what follows. Suffice it to say that, for the purposes of this thesis, some of the main theses of the Standard Model of social ontology are embraced, mainly in the form of Searle's theory, in order to describe legal reality and legal concepts as socio-institutional categories. This latter theoretical initiative has already been launched by the so-called (neo)institutional theory of law.

1.2. Social ontology and law: neo-institutionalism

The institutional theory of Neil MacCormick and Ota Weinberger, and more lately Dick Ruiters, are some of the best-known applications of a social ontology approach, often in a Searlian fashion, to legal theory (MacCormick and Weinberger 1986). These authors are generally ascribed to the strand of 'new' institutionalism (here also neo-institutionalism), in contrast to the classical strand of institutionalism whose main exponents were Maurice Hauriou (1856-1929) and Santi Romano (1875-1947).

Both forms of institutionalism arise within the context of critical theories of legal formalism and, in general, of visions of law as a logical and systematic science. Despite its legal positivist matrix,¹⁰⁷ institutionalism differs from it in that it takes more account of the social sphere, and its normative dimension, in the way it hosts and determines the legal realm. This inclination to look at law as a social phenomenon is undoubtedly one of the elements of greatest continuity between the various forms of institutionalism.¹⁰⁸

Roughly, institutional theories seek to describe law and its ontology in terms of an institutional (normative) order, even though they have different ideas about what 'institutional' actually means. Classical institutionalism appears more deferential to the sociological tradition,

¹⁰⁷ While Santi Romano's institutionalism is distinctly legal positivist, the same cannot be said for Maurice Hauriou's, which is more closely linked to natural law ideas (namely, Thomism). See (La Torre 2016).

¹⁰⁸ This is not the only aspect; they are also united by ideas such as, for example, the separation of law and morality and the central role assigned to language. On this point, see (La Torre 2010, 110).

which attributes to institutions the meaning of social structures and organizations. In this context, Hauriou tends to think of institutions as forms of social aggregates that precede and produce law, and which law would be tasked with holding up. Romano, instead, combines the concepts of institution and legal order, in the sense that every legal order is seen as a social institution and, vice versa, every social institution is seen as a legal order as it is internally organized by means of rules and roles (Croce 2012, 112).

Neo-institutionalists for their part do not employ a generically social notion of institution to describe law and tend to reject the identification of social institutions with legal orders. There are two senses, they claim, by which law is an institutional phenomenon: in a sociological sense “in that it is in various ways made, sustained, enforced and elaborated by an interacting set of social institutions” and in a philosophical sense for which “the existence of a valid rule of law, as of a valid contract, is a matter of institutional fact” (MacCormick and Weinberger 1986, 56). Therefore, in the wake of Anscombe and Searle, the neo-institutionalism of MacCormick and Weinberger is based, even before the notion of institution, on that of institutional facts. They believe that law belongs to the latter category of facts, as opposed to brute (empirical) facts, and that institutions are general categories or abstract concepts with normative value, which guide the conduct and interactions between individuals.

In Weinberger, institutions seem to merge into norms, as the former have at their core an element of practical information, i.e., the norm: “The institutions are realities located in the context of action; they constitute models of action, spheres of possible action, and organize interaction among men.” (Weinberger 1986, 117).

Slightly different is MacCormick's reading for whom institutions appear more as the conceptual background of institutional facts, in a manner that is sometimes not well defined: “Lurking in some Platonic cave behind the institutional fact lies the institution itself. Searle tell us that institutions are systems of rules, indeed, in his very own words ‘systems of constitutive rules’. But that [...] would simply involve an obvious confusion between the law of contract and the legal institution ‘contract’ itself which is regulated by that branch of law. Institutions (and institutional facts) in the philosophical sense obviously have something to do with rules, but are not identical with them. If we want to make clear this philosophical notion of an institution we shall, I think, do well

contemplate [...] the one which Buckland had in mind when he called his book *The Main Institutions of Roman Private Law* or Renner when he called his book *The Institutions of Private Law and their Social Functions*.” (MacCormick and Weinberger 1986, 51).

The preference for the latter cited literature is typical neo-institutionalist move as, for instance, Karl Renner aims to combine a strictly legal analysis of institutions with one that pertains to the economic and social functions that those institutions perform. “Every legal institution” Renner says “is to a conceptual approach a composite of norms, a total of imperatives”, that “regulates the factual relationships of living beings and successive generations, [...] facts which are in a constant state of flux and it [legal institution] is – like law in general – nothing but one aspect of the subject-matter which it governs” (Renner 1949, 53).

Later, MacCormick will distinguish three meanings with which the notion of institution can be used in law: (a) institution-agencies, (b) institution-arrangements and (c) institution-objects. The first expression (a) typically designates Courts, government departments, cabinets, and even companies or corporations with legal personality. The second (b), on the other hand, designates contracts, marriages, ownership, succession, trusts, adjudication, family, personality, and other agreements between persons and/or legal entities. Finally, institutions-things, (c), refer to non-tangible objects that are created by legal provisions, e.g. copyrights or shares (MacCormick 2007, 36). For the sake of this thesis, the sense of institutions of interest is the second, (b), which refers to systematic legal concepts, which include legal personality, a topic MacCormick has devoted special reflections to (MacCormick 1988).

Anyway, these three senses have in common the fact of being particular type of arrangement of normative provisions and relationships. All of these institutions, or more precisely the instances of legal institutions, are governed by a triad of rules that establish (1) the conditions for the existence of a single instance of the institution – i.e. the existence of *a* contract or *a* person of law (*institutive rules*); (2) what legal consequences follow from the existence of an instance of the institution (*consequential rules*); and (3) the conditions by which individual instances cease to produce their effects and ultimately to exist (*terminative rules*).

Despite the Searlian imprint discernible in this conception of (legal) institutions, MacCormick is not willing to fully embrace the notion of institution as systems of *constitutive* rules, as stated in the passage quoted above and proved by the triad of rules he says constitute institutions, which looks like a combination of both constitutive and regulatory rules, i.e. “Regulative rules regulate preexisting forms of behavior, constitutive rules make possible new forms of behavior” (Searle 2018, 51). Along these lines, the pair of Searlian rules overlap and intertwine with MacCormick's triad of rules, which can be described as both constitutive and regulatory (MacCormick 1998, 335; Ruiter 2001, 73). We will return to these rules, and the debate about their nature, when discussing the conceptual analysis of legal personality.

To recapitulate, neo-institutionalists see law as consisting of institutions (e.g., property, contracts, and personalities), essentially theoretical terms gathering complex sets of rules by which both to create practices and to guide and interpret a cluster of human conduct. In the words of MacCormick: “The term ‘institutions of law’ [...] is therefore to be understood as signifying those legal concepts which are regulated by sets of institutive, consequential and terminative rules, with the effect that instances of them are properly said to exist over a period of time, from the occurrence of an institutive act or even until the occurrence of a terminative act or event” (MacCormick 1986, 53).

As a result of the implementation of these rules, law is an institutional fact. The advantages of this approach derive from exploring law both as a set of acts and facts comprehensible only in light of appropriate networks of rules, gathered in abstract meanings or institutions-concepts, and as a normative order encompassing a set of social practices.

I believe that the social ontology approach subscribed by MacCormick presents elements of originality that are also useful to grasp for the metaphysical framework of legal personality (as will be better expounded in section 6). Yet, his notion of institution-concept suffers from a certain ambiguity or incompleteness probably caused by an excessively dogmatic vision, which emphasizes their role as functional devices for the organization of legal thought while neglecting their ontological status (MacCormick 1992; Latorre 1999). Such a view, i.e. of institutions as mere conceptual boxes reorganising more fundamental material, lends itself to a number of criticisms. Indeed, MacCormick himself does not seem to take this argument all the way, and lays the

groundwork, not fully developed, for a different line of reasoning, as signalled by the frequent reference to the 'final causes' or 'points' of institutions. We will deal with these profiles in more detail in the next chapter in which two readings of legal concepts will be contrasted: inferentialism versus ontologies. For the time being, it is enough to point out that MacCormick's account of legal institutions is a good starting point to be integrated with some insights coming from the contemporary debate of philosophy and social sciences.

In the next section, I will try to give a more complete account of the notion of institution by virtue of the contemporary debate of social ontology about the nature of human institutions, trying to apply it to legal institutions.

1.3. What institutions are? The rules-in-equilibrium account

MacCormick somehow adheres to a social science tradition that characterises legal institutions as systems of rules that guide the conduct and interactions of human beings. One of the most prominent formulations of this paradigm, for institutions in general, is that of the economist Douglass North: "Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction. In consequence they structure incentives in human exchange, whether political, social, or economic. Institutional change shapes the way societies evolve through time and hence is the key to understanding historical change [...] They are a guide to human interaction, so that when we wish to greet friends on the street, drive an automobile, buy oranges, borrow money, form a business, bury our dead, or whatever, we know (or can learn easily) how to perform these tasks" (North 1990, 3). From this perspective, institutions contain both formal rules, laid down in laws, constitutions or regulations, and informal rules, originating in conventions and non-legal norms.¹⁰⁹

A famous version of this rule-based approach is, as already anticipated, Searle's, albeit specified with regard to the quality that these

¹⁰⁹ A hierarchy of the rules of which institutions are composed is elaborated by Elinor Ostrom, who distinguishes between (1) *operational rules*, which govern ordinary interactions, (2) *collective choice rules*, which serve to select the ordinary ones, (3) *constitutional rules*, which are used to select the ones of collective choice, (4) *meta-constitutional rules* and finally, as a creative constraint, the (5) *biophysical world*. (Ostrom 2005, 58).

rules must have, i.e. they must be constitutive. Yet, MacCormick's theory is perhaps more similar to North's, given that the triad of rules constituting institutions of law (institutive, consequential and terminative) also includes what are typically called regulatory rules, i.e. rules that regulate pre-existing facts.

In any case, this approach sees institutions as systems of rules that serve to facilitate human interactions also through the attribution of powers and duties. Thus, in the words of Hindriks and Guala, marriage is technically an institution because: “Married couples have rights and obligations that indicate what they must and must not do when they engage in certain activities. In most Western countries both husband and wife are responsible for procuring the material resources to support their family, for example. They are both responsible for their kids’ welfare and education; they have a mutual right of sexual monopoly [...]. The reason why such rules exist is fairly obvious: they help husband and wife attain goals that would be more difficult to accomplish if they acted independently, in an uncoordinated manner” (Hindriks and Guala 2015, 462).

This point of view is particularly suited to economic institutions, which are studied with regard to their effects and incentives on human interactions and other phenomena, as is the case, for example, when trying to study the effect of economic institutions – e.g. the single currency, accountancy rules, private ownership or competitive markets – on economic growth (Hayek 1973; Acemoglu 2003).

The rules account also seems at first glance adequate when applied to legal institutions. Legal personality itself is effectively described by this account, also with regard to the facilitating effect on coordination and human interactions. This second profile emerges especially from a perspective of economic analysis of law which, as noted in the first part of this thesis, pays attention to the incentives or distributive effects of legal personality, for example in the form of the overall effects that the attribution of personality to collective entities – e.g. companies – may have on economic growth (Hansmann, Kraakman and Squire 2005; Huff 2003).

However, on closer inspection, this approach does not seem to exhaustively explain what institutions – social or legal – are or do, because in a number of circumstances these institutions are not strictly observed as rules. In some cases because rules are simply not enforced, e.g., the prohibition, never enforced, for women to wear trousers in a

French law of 1799, or also, albeit at a different level, the rule stated in art. 39 of the Italian Constitution that recognises legal personality to trade unions and that has never been applied (at the behest of the trade unions themselves). In other cases because the rules, although universally recognised as valid, are exposed to habitual and accepted deviations, this is the case of the motorway speed limit set at 65 mph, but whose application is subject to a margin of tolerance up to 75 mph (Greif and Kingston 2011). This deviation gives drivers an incentive not to exceed the 75 mph limit, and the police have an incentive not to fine those who stick to that limit. So this happens to be the case for several legal institutions.

The alternative, equilibria-based view posits that institutions are not reducible to prescriptive rules, but are patterns of behaviour constituted by regularities and shared expectations: “Such regularities ‘can be best described as non-cooperative equilibria’ of strategic games [...] because out-of-equilibrium actions are unstable and are unlikely to be repeated in the course of many interactions.” (Hindriks Guala 2015, 463). In game theory an equilibrium is the best combination of strategies available to the participants of an interaction, such that none of them has any incentive to change strategy unilaterally.

Of course, not all equilibria are institutions. To be an institution an equilibrium must (a) concern a coordination problem and (b) requires players to correlate their strategies. Thus, for example, the equilibrium in a prisoner's dilemma is not an institution because agents can follow the associated strategies of that equilibrium unilaterally, without coordinating with each other (Hindriks Guala 2015, 466).

A classic example for describing institutions as equilibria in non-cooperative games is actually that of a legal institution, i.e. private property. The game of private property, Guala and Hindricks point out, can be depicted as a *correlated equilibrium* of the hawk-dove game. In order to appropriate a plot of land, the interested parties, suppose they are two identical players, have two possible strategies: to behave like hawks, both claiming the land and ending up fighting; or to behave like doves, without fighting, but renouncing the utility of the land appropriation. The best combination of strategies is, intuitively, the one where one of the two players uses land and the other refrains from doing so. Private property can thus solve this coordination problem by means of *correlation devices*: in order to coordinate access to property, players can follow a

precedence or bidding rule, e.g. whoever makes the highest economic bid gets access to the land.

However, an account based only on equilibria would still be incomplete in describing human institutions, as correlation devices are also present in interactions between non-human animals (so called 'animal conventions'). Unlike the latter, in fact, human institutions use representational devices, symbolic markers, in order to condition and coordinate behaviour, or create new patterns of conduct. And it is here that Hindriks and Guala merge the two accounts: "[...] rules are representations in symbolic form of the strategies that ought to be followed in a given game. Just like the rules account without equilibria is incomplete, so is an equilibrium-based account without rules. A satisfactory theory must combine the best features of both." (Hindriks Guala 2015, 467). In a nutshell, as the actions of strategic games can be expressed as rules – which by the way summarize the properties of equilibria – the two accounts are compatible (Aoki 2001).

I believe that Hindriks and Guala's unified theory can also be exploited to describe the peculiarities of legal institutions, especially to visualise them in a multidimensional way, as anchored to a network of behaviours and socio-cultural patterns of a conventional nature and not flattened on the legal dimension alone. The ability of legal institutions to govern interactions and uncertainty not only rigidly – i.e. by means of prescriptions (or constitutions) of behaviour – but flexibly – i.e., by means of informal or interpretative equilibria – is thus emphasised.

This is particularly advantageous if, as in the case discussed in this thesis, we aim to conceptually address a coordination problem (as seen in Part 1, Chapter 3, sec. 6, and further elaborated later on).¹¹⁰

A unified notion of institution as reconstructed in this account is, finally, functional to MacCormick's own neo-institutionalist approach, which aspires to reconcile the sociological and jurisprudential dimensions of legal discourse, with the aim not only of connecting the world of experience with that of legal theory as well, but the opposite direction, i.e. how law influences social conduct: "[...] there remains a somewhat different 'sociological' question not a causal one, but one couched in the terms of rational understanding of social action [...]. It

¹¹⁰ Hindriks and Guala's theory consists of another part in which they try to deconstruct Searle's constitutive rules-based framework, and in which they try to argue that constitutive rules can be derived from regulative ones through the introduction of theoretical terms.

is, after all, people as social and moral agents, who both make and use the law, and sometimes break it. For them law is thinkable essentially as, if not only as, proving grounds for action and for reaction to actions. It is, ad Luhmann says, a basis for expectations of expectations among people. And here lies the real challenge to Kelsen. The question is not how law *causes* behaviour, but how it could intelligibly *motivate* it in the sense of giving reasons for it” (MacCormick 1988, 373).

1.4. Legal institutions as social kinds

In the context of social metaphysics, legal institutions are subspecies of social kinds. In the next section, I will show the metaphysical framework of social kinds, and so legal institutions, in Brian Epstein’s account. However, Epstein’s notion of social kinds is not particularly detailed: “[...] it is useful to think of them as the categories we might use in the social sciences, but remain open-minded about the sorts of categories these might be. [...] Social kinds—like social properties—have fixed instantiation conditions. Or, more appropriately, we might say that kinds have fixed membership conditions. The conditions under which something is a member of a social kind are the same across all times and possibilities” (Epstein 2015, 68). To better capture the peculiarities of legal institutions, I will follow a more detailed account of social kinds.

Kinds that make up social reality are, in the traditional sense (e.g. Searle), those things that can only exist on account of human activity, because of our propositional attitudes towards them: e.g., money, borders, corporations, the prime minister; it follows that they are ontologically subjective (though being epistemically objective) (J. Searle 1996). Hence, for something to count as money, according to Searle’s paradigmatic example, it suffices to be regarded and treated as such by virtue of our collective conventions, beliefs and attitudes.

Although this view is convincing in many respects, it fails to recognize that some social kinds do not meet these requirements, as Amie Thomasson has brilliantly argued: “The possibility is often overlooked, however, that there may be entities (and kinds of entities) depending on mental states of various kinds, without their depending on any beliefs about them (or about that kind itself). Some social kinds such as racism, superstition, etc., do depend on the existence of certain sets of beliefs and intentional behaviours, but may exist without the existence

of any beliefs that are themselves about racism, superstition, etc” (Thomasson 2003, 606). As it is the case of recession, since a state can be in recession, even if nobody notices it or considers it as such.

Searle accepts Thomasson's observation in part, acknowledging that there may be social kinds that do not require propositional attitudes toward them, though he does not include kinds like recession among the actual social (institutional) facts, but rather as consequences of institutional facts that do not carry deontic powers, or also as *systematic fallouts* (J. Searle 2010).

To specify which genre of social kinds legal institutions are, we may consider a recent classification of social kinds, developed by Muhammad Ali Khalidi, which can include legal institutions too (Khalidi 2013; Khalidi 2015; Khalidi 2016 ; Khalidi 2018).

Like Epstein, Khalidi sees social kinds as analogous to natural kinds, i.e. as sets of proprieties in the social domain. In natural kinds, these properties are linked through causal connections and, for this reason, these kinds can be conceived as nodes within causal networks: “Whenever one property, or a number of regularly co-occurring properties, reliably causes one or more other properties, we have a good candidate for a natural kind” (Khalidi 2019, 2). Yet, causal networks can occur both in the natural and in the social domain.

Differences between natural and social kinds are, for example, that the latter are somehow dependent on humans and their mind,¹¹¹ are interactive – in the sense that they can change as our beliefs change, and do so in return, triggering a looping effect – some of them, as suggested by Searle, depend on our propositional attitudes and, finally, they have a normative dimension (Khalidi 2013).

Khalidi distinguishes three different sets of social kinds: (a) those whose existence, either in general (*type*), or as an individual instance of it (*token*), does not depend on any propositional attitude of humans towards them (e.g. recession and racism); (b) those whose general-type existence depends on particular attitudes of humans towards them, although these attitudes need not occur towards each of their individual instances for them to be token of those types (e.g. money and war); (c) those whose existence as either a type or a token both depend on the

¹¹¹ All social kinds are mind-dependent, even though according to different degrees, as also shown by the differences between the following three sets of social kinds.

propositional attitudes humans have towards them (e.g. prime minister and permanent resident) (Khalidi 2015).

The first category thus includes those kinds which, being social, arise in the context of human interactions, but which can also exist without any specific propositional attitude towards them (weak mind-dependence): for example, for *racism* to exist in a society does not require every member of that society, whether victim or perpetrator, to be aware of it, either as a category or as a concept. For this latter reason Khalidi claims that these social kinds are only not mind-independent but also concept-independent.

The second set of social kinds exists by virtue of some sort of type-recognition, without being constrained to any token-recognition: for example, for there to be a *war* there needs to be a concept of war and a declaration between the parties, as well as other practices associated with this concept, but not every individual act of war to be so needs recognition by the parties. These type-kinds are therefore both mind-dependent and concept-dependent, but not all tokens need to be so.

Finally, the third set of social kinds, which includes according to Khalidi the more strictly institutional or conventional kinds, poses the most stringent conditions of existence of the three: for something to count as, say, *permanent residence*, not only are attitudes towards its type required, but it is also necessary that at least some members of society have particular attitudes towards it as a token, perhaps verifying that certain conditions are individually met. Every token of this kind must then be both mind-dependent and concept-dependent: “[...] no one could be a permanent resident of a certain jurisdiction without being recognized as such by at least one government official. In other words, the concept must be applied to each individual member of the kind” (Khalidi 2019, 6).

It should be stressed that while the nature of the first and second sets of social kinds depends partly on their causal properties, as they put physical or causal constraints on their tokens – e.g. ice cubes cannot count as money – the nature of third, both for what concerns types and tokens, depends directly on human mental states. In particular: “[...] the properties associated with them tend to be explicitly stated in a set of rules or conventions. Therefore, if such a kind is associated with a set of properties, that is not because there are causal connections between these properties or because they are linked together by laws or empirical generalizations. Rather, it is because a social institution or community

has decided to associate these properties with the kind” (Khalidi 2015, 106).

Under this theory of social kinds, it is not uncontroversial whether legal institutions, say legal kinds, fall within the second or the third set. Law as a phenomenon in general might fall into the second category, but this might not apply to all legal kinds, as Khalidi acknowledges: “[...] even though law itself is not token-concept-dependent, other legal kinds do seem to be such that each instance is concept-dependent. For example, [...] in many legal systems, a felon cannot be such unless convicted in court. Similarly, a particular jury must be recognized as such to be a jury. Thus, law and other legal kinds are all concept-dependent, while some legal kinds are such that their instances are also concept-dependent” (Khalidi 2019, 7).

It seems to me that most legal kinds fall into the third class, especially taking into account two clarifications that Khalidi offers later. Indeed, even if social kinds of the third class merely result from conventions, or legislative fiat, and not from causal connections between the properties involved, it is not a priori excluded that they somehow participate in causal connections. They can do so in two ways: because the associated properties, and membership conditions in the kind, have been formalised in accordance with causal models pre-existing the rules or regulations (*caveat 1*); because properties associated with the kind by convention or law can actually *create* causal connections, or participate in new causal patterns, that did not exist before the conventional kind was created (*caveat 2*).

Where does the legal kind of personality fit into this framework? I believe that there may be alternative views that are all consistent with the preferred legal-philosophical standpoint. Yet, I think the distinction between different meanings of personhood described in section 1.1. may be useful here. In fact, it seems to me that legal personality in its instrumental sense, as a mere point of imputation of legal positions, can easily be placed in the third class of social kinds, just think of a company or a corporation: “[...] a corporation could not be a corporation without being explicitly conceptualized as one, at least by the owners or shareholders. In these cases, not only does the kind as a whole need to be recognized under the relevant concept, each instance of the kind needs to be so recognized” (Khalidi 2019, 6).

In the two other meanings of person, i.e. co-extensive with human person or intentional agent, this consideration may not hold, especially

if one assumes that the status of person is associated with certain natural traits and capacities.¹¹² On the other hand, Khalidi's caveats may be used to neutralise these essentialist/realist arguments. Especially by virtue of *caveat 1*: i.e., the way in which the status of legal personality may appear to be associated with causally related properties is justifiable on the basis that its formalisation occurred when certain social roles and prerogatives were already in place. Hence, I would tend to collocate personality of law it into the third category of social kinds.

I will come back on this issues when addressing the theories of legal personality (sec. 5 and 6).

2. Analytic metaphysics and the structure of legal kinds

A relevant theoretical debate in contemporary social ontology, which is also relevant for the purpose of providing a description of the structure of legal kinds, concerns the use of notions of analytic metaphysics to describe the building blocks and building relationships of social reality and its constituent parts: social kinds and facts.

Here I should specify that I assume legal reality to be a subtype of social reality, which means that law is made up of facts and objects that could not take place without some social recognition and interaction. As we have seen, one way of describing the social matrix of law is through recourse to the notion of (social) institution, whereby legal facts depend both on systems of rules and on acts and expectations (rules-in-equilibrium account). In addition, these institutions are generally conceivable as mind- and concept-dependent social kinds. Therefore, explaining the nature and structure of law is ultimately an exercise in social metaphysics.

One of the most innovative uses of tools from analytic metaphysics to describe social reality is the one advocated by Brian Epstein in 'The Ant Trap - Rebuilding the Foundations of the Social Sciences' (2015).

Epstein describes the content of social reality through two theoretical categories: social kinds and social facts. Social kinds are used to refer to

¹¹² Recently, the placement of legal institutions in the third class of Khalidi's social kinds has also been emphasised by Paweł Banaś: "[...] legal institutions, including legal personhood, are to be located in the third category, for which collective recognition is not only necessary for the very kind existence but also for that kind membership", in (Banaś 2021, 44).

everything that is part of the social inventory: (a) artefacts such as money, borders, works of art or religious symbols; (b) associative entities such as corporations, institutions, social groups and classes, religions and nations; (c) statuses, roles and social properties including sexual gender, race, private property, public office and family roles. Social facts, however, are different. For a certain philosophical tradition, a fact in general is the referent that makes a proposition true; or, according to an alternative view, a fact is itself a true proposition. In this context, a social fact can be seen as a proposition, or the referent of a proposition, which has as its constituent any social kind. For example, the proposition 'twenty-two individuals compete in an athletic competition' is a different fact from 'Inter and Milan play football', because the latter proposition has as constituent a social kind: football (team).

Epstein addresses some criticisms to the current formulations of ontological individualism – the thesis according to which social facts (and kinds) are exhaustively constituted by the facts about individuals and their interactions – which he considers "trapped" by an excessive anthropocentrism. His aim is actually broader: to reformulate the social ontology debate in the light of the intuition that the two traditional enquiries – into the origin and essential properties of social entities – occur at different levels and moments, and are observable through two distinct metaphysical relationships: *grounding* and *anchoring*. Since social facts and social kinds are, according to Epstein, the result of both relationships, and since the two metaphysical investigations are not *prima facie* contradictory, he proposes a model that integrates them in a coherent way: "Any given social fact has building blocks, and also metaphysical reasons for why that fact's building blocks are what they are" (Epstein 2015, 74).

Epstein finds Searle's account of social reality functional in showing the possible coexistence of both investigations. The constitutive rules of an institutional fact establish the conditions that must be fulfilled for someone or something to count as a certain kind or social fact. Take as an example the fact that 'the banknote in my wallet is a euro banknote', preceded by the following constitutive rule:

RC: Every banknote (z) validly issued by the National Central Banks (X) counts as a euro in the countries that are part of the European Monetary Union (Y).

This constitutive rule stipulates that in order for something (z) to be legal tender euro (Y), specific antecedent conditions must be fulfilled (X). Which implies that the antecedent conditions are in some metaphysical *constitutive* relation of the social kind or fact. To describe this relation Epstein uses the first of the diagnostic tools characterising his model, i.e. grounding: a metaphysical relation of non-causal, synchronic, asymmetrical, non-monotonic and unreflective dependence between *priority* facts – grounds or grounding facts – and *secondary* facts – grounded facts (Schaffer 2009; Rosen 2010; Fine 2012; Audi 2012).

The grounding relation assumes that social (and non-social) reality is hierarchically ordered and that questions about its structure can be explained by relations of existential dependence between facts or kinds. The explanatory intent of the grounding relation is moreover signalled by its asymmetrical nature: the grounded fact exists *because* the grounding fact exists and not vice versa, implying that it is metaphysically necessary that if the latter is obtained then the former is also obtained.

The fact that 'the banknote in my wallet has been validly issued by the central bank' (1.1) *grounds* the fact that 'the banknote in my wallet is a euro banknote' (1.2) does not mean that (1.1) is the causal reason for (1.2), but that it is its metaphysical reason, i.e. that (1.2) is obtained because (1.1) is obtained, and not vice versa. Grounding is therefore, as anticipated, a different relation from causation: e.g., the fact that there is legal tender may be *caused*, among other things, by the need to simplify transactions, but it is *constituted* by the fact that a banknote/currency is issued by a central bank.

There are different forms of grounding. For example, the fact that '22 football players play at the San Siro' (2.1) is a *partial grounding* of the fact that 'the derby is played at the San Siro' (2.2) because (2.1) is not the sufficient metaphysical reason to explain (2.2), but contributes to it together with other facts. If, on the other hand, a grounding fact is the metaphysically sufficient to explain the grounded fact, then there will be *full grounding*: e.g. 'The fact x maxims happiness' totally grounds 'x is right' (utilitarian view).

Usually social facts and kinds are grounded by a heterogeneous set of grounds. Here the Searlian account is still useful, as Epstein tends to see the constitutive rules à la Searle as the rules that group together this set of grounds, the grounding conditions of a social fact or kind; though the rules themselves are not among the grounding conditions.

The author decides, however, to diverge from the logical form of the Searlian constitutive rule, 'X counts as Y in C', which is replaced by the second cornerstone of the GAF model, the *framing* relation, which results from the implementation of a *frame principle* (FP), of the type 'For all z, the fact z is X grounds the fact z is Y':

FP: For all z, the fact that z is a banknote validly issued by the National Central Bank *grounds* the fact that z is a euro banknote in European Monetary Union countries (Epstein 2015, 78)

More specifically, Epstein introduces the *frame*, i.e. a universe of possible worlds sharing the same grounding conditions (of a social fact or kind) as fixed by the same rule, i.e. the frame principle. This frame encompasses both the actual world and the set of potential situations in which, under varying grounding facts, something counts as a euro banknote as they are governed by the same frame principle. Frame principles do not have a stable logical form and, at first glance, might seem to be generalisations of constitutive rules. This is not so: Epstein aims at correcting a structural defect in the notion of Searlian constitutive rules. He believes that these rules confuse the conditions of *existence* and the conditions of *constitution* of social entities.¹¹³

The final step of the GAF is to investigate the origins of the grounds of a social fact or type, i.e. the reasons why there are certain specific grounding conditions rather than others. In short, the relationship between certain facts and frame principles (which these facts bring into being). Epstein believes that this relationship, which he calls *anchoring*, acts as a glue of both social and natural kinds and facts (Epstein 2015).

Returning to the example of money, we can distinguish the two metaphysical relations: the fact that 'the banknote in my wallet (z) is a euro banknote (Y)' is *grounded* in the fact that 'the banknote in my wallet

¹¹³ More in detail, according to Epstein, if we apply the rule in the Searlian logical form 'X counts as Y in C', the question whether a euro banknote exists at time t is equivalent to asking whether the conditions for a piece of paper (X) to count as a euro (Y) in C are satisfied at time t. This for Epstein is implausible since the grounding conditions for the existence and the grounding conditions for the constitution of a social entity are determined differently: e.g., the fact that a euro banknote exists at t may necessarily be grounded by the fact that someone authorised the issuance of the first euro banknote; but such a fact does not determine the constitution of the euro banknote at t, i.e. the contingent fact for which that piece of paper counts as a euro banknote (Epstein 2015, 161).

is a banknote validly issued by the National Central Bank (X)'; the latter fact is in turn *anchored* in the fact that European legislators were engaged in the socio-legal practice of agreeing on rule X of the European monetary system.

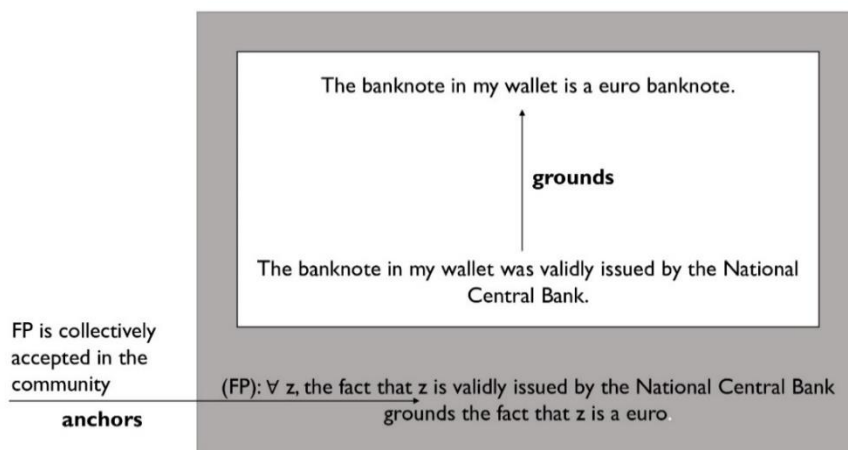


Figure 4. GAF applied to the money case

In the case shown above there is only one grounding condition, but it is usually a heterogeneous set of facts that helps to ground most social kinds. Yet, the mechanism remains unchanged and aims to bring out two types of questions: firstly, what facts and how a social kind/fact is grounded (*grounding project*); secondly, what makes it so that it is precisely those facts that ground one social kind/fact rather than others (*anchoring project*).

The existence of differentiated metaphysical projects facilitates that separation between the investigation of the origin of the social world and its constitution, and from which ontological individualism – as well as social ontology in general – should benefit in order to clarify the terms of the different issues, and avoid promoting anthropocentric theories that make every social fact totally dependent on individual facts.

According to Epstein, the GAF model would be ecumenical with respect to the specific visions of the social world, and indeed would allow for clarity even within theoretical positions, as is the case with

individualism, which in this perspective can be broken down into grounds individualism and anchors individualism. Classical ontological individualism is then to be regarded as individualism on grounds. Individualism on the anchors, on the other hand, tends to see the frame principles as being anchored only in facts about individuals, without committing itself to the same thesis for the grounds, as exemplified by Searle's theory: "Searle is an individualist about anchors, but not about grounds. The facts that ensure that Billy is money in the United States are not facts about individuals – they are not even facts about human beings, actually. Human beings and their intentions only play a role at the level of anchors: they are what makes facts like being issued by the BEP [Bureau of Engraving and Printing, n.d.a.] the ground for being money in the United States" (Guala 2016, 137).

2.1. Grounding-anchoring-framing (GAF) applied to law

The case study Epstein considers in order to test the GAF model is law (and its nature). In particular, the observation angle of the legal phenomenon is a specific conception, the theory of H. L. A. Hart laid out in 'The Concept of Law' (1961).

Notoriously, Hart maintains that the legal system is a dynamic normative system composed of two types of rules: secondary and primary rules. According to Hart, secondary rules are power-conferring rules, i.e. those that identify, modify or enforce primary rules, which in turn prescribe or prohibit certain conducts to individuals. Secondary rules are thus norms above norms, or meta-norms. The parameter for decreeing the validity of norms, i.e. their belonging to the legal system, is entrusted to a specific secondary rule, socially practised by legal operators: the rule of recognition (RR).

In order for the rule of recognition to correctly fulfil this function, at least two conditions must be satisfied: (1) that legal operators follow a social practice of observance of RR; (2) that this practice is based on an attitude of acceptance of RR. It follows that the rule of recognition is a social rule accepted by legal officials as a common standard of behaviour with binding power, and is a condition of existence of the primary rules, as well as of the legal system in general.

Epstein believes that the legal system, as conceived by Hart, can be reproduced and explained through the grounding-anchoring diagram and that legal norms can be equated with frame principles.

Consider a legal fact such as 'Luke Skywalker is guilty of graft' (3.1) [extortion by a public official]. This fact depends on overriding facts, some of which are Luke Skywalker's individual facts: e.g. his role, concrete behaviour, mental states, etc. Other facts relate to what makes someone legally guilty of graft. The situations in which this legal fact occurs are laid down by law, in Italy by Article 317 of the Criminal Code: "A public official or a person in charge of a public service who, abusing his position or powers, compels someone to give or promise unduly, to him or to a third party, money or other benefits, shall be punished by imprisonment from six to twelve years".

In this sense, legal rules set out the grounding conditions of the legal fact. The fact that 'Luke Skywalker is a public official' (3.2) together with the fact that ' Luke Skywalker compels someone to give or promise unduly...' (3.3) and...(3n), ground the fact that 'Luke Skywalker is guilty of graft (3.1). In other words, (3.1) obtains because (3.2) and (3.3) are obtained.

So, the rules of the criminal code presented here are primary rules that govern legal reality, just as the framing principles govern the set of possible worlds that share certain basic conditions; however, they do not cover the legal phenomenon, which also depends on the wider social practice that brings it into being. This suggests that the legal fact (3.1) is grounded in the facts described by the primary rules, i.e. the framework principles, but is also anchored in the facts that put the primary rules themselves into being. The correct functioning of this recognition rule is in turn conditioned by the presence of practices, beliefs and attitudes towards the rule itself. In other words, certain facts – which for Hart are conformity and acceptance – bring the social rule of recognition into being, i.e. they *anchor* it. What results from all these "hartian" relationships is schematised by Epstein as follows:

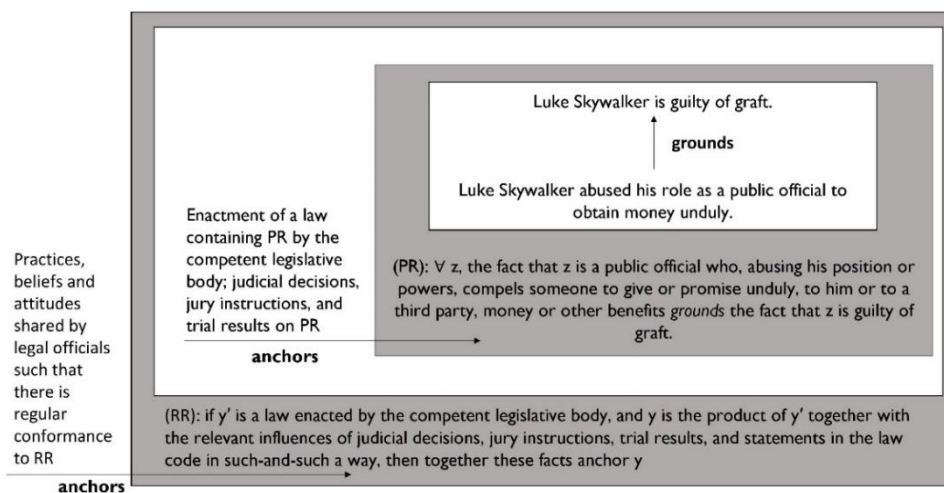


Figure 5 - GAF applied to Hartian conception of law

As can be seen, in contrast to Figure 1, two anchoring relationships appear in this diagram: one forming the basis of the frame principle that sets the grounding conditions, i.e., the primary rule (PR), and the other forming the basis of the frame principle that sets the anchoring conditions, i.e., the recognition rule (RR). The two sets of facts acting as anchors are qualitatively different, the more external one being composed of practices – “convergent behaviours and attitudes towards those behaviours” (Epstein 2015, 97)– which anchor the rule of recognition fixing criteria of legal validity stated in a primary rule of a context (the frame), which in turn is anchored to the more internal set of facts consisting substantially of procedures.

This model, which is useful for studying the building blocks and origins of social types and facts, including legal ones, can of course be extended to conceptions of the nature of law other than Hartian ones.

2.2. A metaphysical framework for legal kinds (and facts)

It should be mentioned that the view based on the distinction between grounding, framing and anchoring (GAF) advocated by Epstein can be contrasted with the more orthodox view of the so called ‘conjunctivism’. A specific conjunctivist critique of the GAF model is that of Jonathan Schaffer (Schaffer 2019).

Jonathan Schaffer's conjunctivist view – the so-called grounding-only (GO) model – is rooted in the idea that social facts and kinds are dependent on facts acting at two different levels: (a) *rule-setting facts* and (b) *move-making facts*. It follows that Epstein's thesis that there are two metaphysically distinct planes in the construction of social entities is preserved by the GO model: i.e. rule-setting facts would correspond to anchors, while move-making facts would correspond to grounds.

However, unlike Epstein, in Schaffer's model the metaphysical relations that build social reality are all grounding relations, while the notions of anchoring and framing have no theoretical place. In other words, although established between different metaphysical planes, the relationships between social rules (frame principles) and facts that bring them into being, as well as those between facts (or kinds) and what constitutes them, are for Schaffer all grounding relationships: the first of the so-called *structuring grounding* and the second of the so-called *triggering grounding*.

To dwell on the entire debate between Epstein and Schaffer would take too long, but for the purposes of this analysis – which is to use tools of social ontology and analytic metaphysics to understand the nature and the building blocks of legal personality – I think it might be functional to derive what these two theories have in common.

First of all, both models use the grounding relation to describe the kind of metaphysical dependence between social (and non-social) facts. In general, reality observed through the prism of metaphysical grounding is hierarchically ordered according to criteria of ontological priority: there are facts that are more fundamental than others, from which the latter are obtained in a non-causal manner. In this sense, grounding can be used, as we have seen, to describe the metaphysical structure of law, that is, the way law depends on or is determined by other facts. Whether or not law is seen as a strictly social phenomenon is not decisive for the application of this model, since in both cases it is plausible that legal facts or types are secondary and derive from more fundamental facts or types, e.g., beliefs, social practices, conventions, natural facts, mental facts, and so on.

If more fundamental facts are the metaphysical reasons for legal facts (or kinds) – the latter exist *because* the former exist – then the grounding relation can effectively describe the origin and the building blocks of

legal reality.¹¹⁴ Thus, leaving aside different philosophical views on the determinants of law, the inquiry into what grounds law is an inquiry into the synchronic and constitutive relations between legal facts and other more basic facts – e.g., facts about the mental states of legal officials – and not into the set of causes for which the coming into being of a legal fact (or kind) is a mere effect.

In the second place, both the model defended by Epstein and the conjunctivist one discriminate two metaphysical levels to which correspond facts with different functions: (a) facts that set up rules, and thus fixing grounding conditions, and (b) facts that constitute social facts and kinds. Although the authors have differing ideas about the quality of the metaphysical relations established at these two levels, both think of them as relations between sets of facts that are not mutually connected.

In Epstein's model, a frame of possible worlds expresses the general grounding conditions through a framing principle, which connects the grounding facts to the grounded facts, and which is in turn anchored to a set of second-level facts.

Schaffer, on the other hand, simplifies the network of relations, using only grounding, while preserving the separation between first-level facts (i.e., move-making) and second-level facts (i.e., rule-setting): the latter *structurally ground* the social rule; the social rule, in turn, grounds, so as to *trigger*, the social fact, together with the first-level facts that it sets.

To conclude, the separation between metaphysical levels underlying the constitution of social kinds or facts may prove fruitful for the study of the structure of legal kinds (or facts): e.g. the legal institution of marriage will be grounded by a set of facts/conditions fixed by legal rules, e.g. the civil code, but also determined at a second, deeper level by a set of facts and kinds which are not fixed by legal rules but which depend on the fact that legal practice is part of a wider social practice: in the case of marriage, kinds and facts, not strictly legal, such as 'monogamy', may affect the way civil code rules, and then legal kinds, are designed. These second-level facts – which can be connected with either an *anchoring* or a *structuring grounding* relationship – make the grounding conditions the specific ones set by legal rules, and not others.

¹¹⁴ That the study of legal ontology centered on the metaphysical foundation can be a fruitful direction for legal theory is evidenced, moreover, by the numerous scientific production of the last years. See, above all, (Rosen 2010; Plunkett 2012; Chilovi and Pavlakos 2019).

2.3. The value of the grounding-anchoring diagram for legal ontology

In this section I defend the epistemic and pragmatic value of the modal distinction between grounding and anchoring, with specific application to the ontology of law: the distinction reveals how we think and use legal kinds and facts.

We know that the argument for the modal distinction is underpinned by the universality thesis: social rules provide universal tools - kinds and facts - that can be exported across situations, times and places where the facts that bring the rules themselves into being, i.e., the anchors, do not exist. Whereas in no case could we obtain a social fact without the constituent facts, i.e. the grounds.

In this context, grounding and anchoring can be compared to the concepts of *etic* and *emic* in anthropology, where by the former we mean the study of a culture through the categories, beliefs and values of another culture (typically that of the observer); by *emic*, on the other hand, we mean the study of a culture through its own categories, beliefs and values (an internal point of view, also referred to as the 'native's perspective'). Therefore, ethical analysis, in order to be conducted, requires cultural kinds and facts to be exported without the relevant anchors (Epstein 2019, 771).

Where, on the other hand, does the epistemic and pragmatic value of the grounding-anchoring diagram lie for law? Two situations need to be distinguished.

At a theoretical-general level, epistemic value is reflected in the way differences between legal-philosophical theories are framed (Epstein 2015: 98). In particular, the debate on the nature of law can be conceived as a debate on the grounds (grounds) of primary rules: e.g., are rules commands or are they reasons for action?

The debate on sources, on validity in law, or on the role of morality, can be conceived instead as a debate on the anchors of primary rules and secondary rules. For exclusive legal positivism, both rules are anchored in strictly social facts, because neither the existence nor the content of the primary rules would require moral evaluations (the so-called 'social sources thesis' applies). For inclusive legal positivism, on the other hand, this would only apply to secondary rules, since even if the existence of a rule in a legal system would always depend on a social fact (the so-called social facts thesis applies), moral reasons could determine the

assessment of the validity of primary rules. For inclusive legal positivism, on the other hand, this would only apply to secondary rules as, even though the existence of a rule in a legal system always depends on a social fact (social facts thesis applies), moral reasons can determine the judgement of the validity of primary rules; the latter are thus anchored to both social and moral facts. Any theory of the criteria of validity of legal rules can be seen as a theory of the facts that anchor primary rules; and any theory of the sources of law as a theory of the facts that anchor secondary rules (given the division between primary and secondary rules).

At a more concrete level, the different modal connectivity of grounding and anchoring reveal the epistemic value in the way legal kinds are formed and applied. We can export, for instance, the status of legal personality to circumstances other than those provided by anchoring facts, e.g., the civil code or the value context in which all human beings enjoy equal dignity. This is what we do when we ask whether Roman slaves had any form of legal personality. Similarly when we consider whether, *de lege lata*, a given entity is a 'de facto' legal person, i.e. without formal recognition, as happens when discussing the legal subjectivity of the unborn child. Or, *de lege ferenda*, with a view to conferring legal status on an entity that previously did not have it, as happens when discussing the personality of environmental entities, non-human animals or artificial intelligence systems.

Finally, in law the distinction also becomes relevant when exporting a legal discipline as designed in a given legal system to a different one, just to assess its convenience and effects. In all these cases, metaphysical grounds are divorced from metaphysical anchors. In short, policy-making activities clearly show that a set of grounding conditions can produce certain legal kinds and facts also with varying anchoring conditions. Of course, this conclusion will only be correct if one accepts the idea that social kinds are real kinds and not nominal ones.

The pragmatic value of the distinction also emerges in relation to the judicial use of legal kinds and facts. For example, the judicial ascertainment of the grounding conditions of a legal fact can be divorced from the ascertainment of the anchoring conditions. This can occur when the ascertainment of the conditions of criminal conduct is deviated from by facts concerning anchors, e.g. jurisdiction.

The distinction can also affect procedural acts and strategies. Epstein himself takes Ratko Mladic, accused of being a war criminal by the

International Criminal Court, as an example. Mladic's defence lawyer could follow two strategies: (a) persuade the judges that the defendant does not fulfil the conditions to be a war criminal, e.g., because he carried out the orders of his superiors or because there was no real war (grounds); (b) question the very conditions required to be a war criminal, e.g., giving certain orders does not count as a war crime (anchors).

For its part, the ontology of law can contribute to enriching Epstein's theory by revealing the opportunity to specify different types of grounds and anchors. Although we know that multiple grounds and anchors contribute to the metaphysical constitution of many social facts, the nature of these facts is not equally specified. A distinction consistent with the characteristics of social, and legal, reality would be that between *procedural* anchors - e.g., legal rules and practices - and *substantive* anchors, which in turn can be distinguished into *axiological* - e.g., attribution of value to human dignity - and *contingent* - e.g., regularities of behaviour and collective attitudes.

Indeed, in Epstein's adaptation of Hartian theory to the GAF, there is an implicit diversification between anchors that link legal kinds and facts at different metaphysical levels. However, the diversification is limited to anchors of primary rules (PR) and anchors of recognition rules (RR), while the configuration of the metaphysical role of the other secondary rules (SR) that are not of recognition is missing: the rules of adjudication and of change. I would suggest in this regard that since the recognition rule is a peculiar secondary rule, which identifies and legitimises both the primary rules and the other secondary rules (Raz 1971), the metaphysical position of the anchors of the other SR is intermediate: they will be anchored by the RR while, in conjunction with the latter, they will anchor the PR.

A perhaps more complex discourse concerns the nature and classification of grounds. A first distinction should be made between normative and non-normative grounds, even if this requires amending a substantial part of the GAF.

Finally, to provide as comprehensive a model of social (and legal) reality as possible, one should consider which facts metaphysically constitute the very anchors of social kinds and facts. Indeed, anchors may themselves be social facts (or may not be) and investigating their metaphysical reasons would help to define their contours and roles.

For this reason I will take Epstein's account of the building blocks and building relations of social kinds/facts as valid in order to advance

a framework – both metaphysical and conceptual – of legal personality.¹¹⁵ This is why I will later try to apply the grounding-anchoring diagram to the relations between layers of the neo-institutional ontology of legal personality we have seen in the previous chapter.

2.4. What about normativity?

While I am persuaded of the correctness of the distinction between metaphysical grounding and anchoring to explain law, I am not so persuaded of the third relation Epstein uses in his model, the so-called *framing*, which relate the social rule, or framing principle, and the grounding fact. I do not think it is clear what metaphysical role this relationship plays, nor how it makes social rules part of the construction of the social world (more on this in the next section). Most importantly, framing and framing principles, as pointed out in the scientific debate, lack the normativity dimension present in Searle's account of constitutive rules (Roversi 2021).

This lack of normativity risks to infect or jeopardise the entire adequacy of the metaphysical grounding relationship used by Epstein to account for institutional and legal kind. Indeed, as Kit Fine argues, normative facts like law and morality are in a peculiar grounding relation – which maintains similar properties of non-causality, asymmetry, transitivity, non-monotonicity, etc. – which is *normative* rather than metaphysical in nature, thus with different and weaker modality: the normative grounds for a normative fact, e.g., a moral fact, is a matter of normative necessity and not of a metaphysical one.

Thus, according to Fine, this normative grounding does not overlap or cannot be defined in terms of the metaphysical grounding: “For consider the fact that a given act was right or not right ($R \vee \neg R$). This is grounded, we may suppose, in the fact that it is right (R). The fact that it is right, we may suppose, is (normatively) explained by the fact that it maximizes happiness ($R0$). So the fact that the given act is right or not right is explained in the generic sense by the fact that it maximizes happiness. But it is a metaphysical necessity that if the act maximizes happiness then the act is right or not right ($\Box(R0 \supset R \vee \neg R)$), since it is

¹¹⁵ I also find Epstein's counter-replies to Schaffer (specifically, to the definition reply and the relations reply) convincing.

a metaphysical necessity that the act is right or not right ($\Box(R \vee \neg R)$). However, the fact that the act maximizes happiness does not metaphysically ground the fact that it is right or not right, contrary to the proposed definition” (Fine 2012, 40).

An alternative view to Fine's, one that seeks to reconcile the two different forms of grounding, is that of Gideon Rosen: although normative grounding is distinct from metaphysical grounding, it would be possible to define the former in terms of the latter. This is made possible by the introduction of a non-naturalist bridge-law, essentially a moral law that contributes to ground normative facts. So normative facts, Rosen argues, are grounded both metaphysically and normatively: to say that G normatively grounds F is to say that G metaphysically grounds F *in conjunction* with a suitable normative law linking G and F (Rosen 2017). To compare the two competing views, let us return to the example of what grounds what is right, for Rosen to say that ‘maximising F's happiness’ *normatively grounds* that ‘F is right’ means that ‘F maximising happiness *metaphysically grounds* that A is right in conjunction with the moral law for which maximising happiness is right’ (Leary 2020, 475).

Rosen believes that moral and legal explanations are equal in this respect, given that legal facts are also grounded in pre-legal facts in conjunction with general rules, e.g. the general fact that it is illegal to engage in a given conduct: “Just as particular legal facts are grounded by subsuming the case under a general legal rule, so particular moral facts are grounded by subsuming the case under a moral principle”. Though with a difference: “The general legal facts that figure in the explanation of particular legal facts are themselves ultimately grounded in the pre-legal facts. If it’s illegal to drive down Main Street at 80 mph, this is so in virtue of some sprawling complex of facts about the sayings and doings of legal officials, pre-legal moral facts, and so on. So even if general legal facts play a role in the grounding of particular facts of the form [A is illegal], the ultimate ground for such facts involves no general legal rule; indeed it involves no legal facts at all” (Rosen 2017, 141).

Rosen's account of normative grounding converges on the distinction drawn by both Epstein and Schaffer between rule-making facts and rule-setting facts, also opening up to the meta-institutional ontological layer; however he seems closer to Schaffer's with respect to the metaphysical role assigned to (legal) rules. In any case, it provides arguments in support of the idea that metaphysical grounding can be adapted to account also for the normative dimension of some facts, at

least from an *epistemic* point of view. The failure to account for the practical normativity of institutional facts could be one of the limitations of a metaphysical model of legal reality based on the grounding relation (Roversi 2021).

3. The metaphysical structure of legal personality

For the sake of clarity, I have so far tried to argue that legal personality can be conceived as a socio-institutional kind or fact: in short, a system of rules expressing groups of non-causally related properties.¹¹⁶ Under these rules, legal kinds – such as contract, marriage, citizenship, and indeed personality – as subspecies of socio-institutional kinds, have a *status function* through which deontic powers, e.g. rights and obligations — are assigned to objects, persons, entities, or events (Searle 1996). Which is to say that the thicker notion of personality, presented above (section 1), plays a role in linking antecedent facts to legal consequences.

However, as already argued, rules do not exhaustively describe institutions: regularities of behaviour are backed up by (systems of) incentives and expectations, hence by equilibria of strategic interactions (Guala 2016). On this account, rules, in addition to their classic functions, act as markers: “Because there are multiple potentially self-enforcing expectations in a given situation, coordination mechanisms, including rules, play an essential role in generating regularities of behaviour and social order. Rules fulfil this coordinating role by specifying patterns of expected behaviour, and also by defining the cognitive categories – signs, symbols, and concepts – on which people condition their behaviour” (Greif and Kingston 2011, 28).

So, to understand how to use the concept of personality, and perhaps legal kinds in general, we need to answer the following questions:

(a) Under what conditions is an entity considered a person in law (*use conditions*)?

¹¹⁶ To be more precise, from the point of view of social ontology, legal personality can be understood as a social property entailing a relationship between a bearer of rights and duties, the legal system, and third parties. Yet, nothing prevents us from speaking of it in terms of a social kind or fact; e.g., ‘Asimov is a person in law in the Japanese legal system.’ This does not change metaphysical relations between facts.

(b) What consequences follow from having personality (*legal implications*)?

(c) What set of facts explains/justifies why the use conditions trigger the legal justifications (*background reasons*)?

By contrast, the thinnest account of personality converges towards, though not matching, a notion that I will elaborate on in the conclusions: legal subjectivity, i.e. the unspecified ability to have or acquire at least one right or duty. From this view, legal subjectivity is not an institutional concept, as it does not specify what rights or duties may be acquired by a legal subject or under what conditions.

Different theories of legal validity and of the source of legal kinds can arise depending on how background reasons and their connection to positive law are understood.¹¹⁷ While (a) and (b) concern the way legal status works pursuant to the law in force, (c) is constantly being debated among jurists and philosophers, especially when different claims arise as to whether or not certain entities should be qualified as persons, as happens with unborn children, nonhuman animals, natural resources, and, in our case, AIs. Of course, the first two aspects are a matter for jurisprudential reflection, too: the normative web around legal personhood is in itself a puzzle over which jurists frequently disagree.

We make use of this analytical framework to approach legal personality. From this perspective being a person in law is a socio-institutional fact, positive law rules specify the ‘use conditions’ and ‘legal implications’ of personality and further factors provide ‘background reasons’ that support such use conditions and legal implications.

Thus, it is important here to disentangle two levels in the analysis of legal personality. The ‘use conditions’ and ‘legal implications’ set forth in legal rules frame the relation between certain facts and legal properties: e.g., the fact that ‘*x is a human being*’ (1) leads to (a1) the fact that ‘*x is a person in law*’ and therefore also to (b1) all implications associated with general legal personality (rights, duties, power, liabilities, etc.). Therefore, the legal status of personality obtains because legal rules exist – e.g., the rules stating that human beings have legal personality,

¹¹⁷ This is Brian Epstein’s insight in his model of metaphysical anchoring relations applied to Hart’s theory: ‘Theories of the criteria of legal validity are theories of the sorts of facts that anchor primary rules. And theories of the sources of legality are theories of the sorts of facts that anchor secondary rules,’ (Epstein 2015).

and that personality has certain implications – that establish the use condition (a) and the legal implications (b).¹¹⁸ We have seen that this noncausal and synchronic relation of priority is generally called ‘metaphysical grounding’. Under this relation, the more fundamental facts *ground* the less fundamental ones, without being causes of them: facts (a1) and implications (b1) obtain *by virtue of* fact (1), and not vice versa (Rosen 2017).

On the other hand, the background reasons (c) for (a) and (b) are a further set of facts, namely, those that support the rules establishing the grounding relation.¹¹⁹ These facts can pertain to legal practice – e.g. enactments in conformity with legal procedure, precedent, and judicial decisions (Epstein 2015, 94) — or they may pertain to the wider practice that law is part of – e.g., social values, habits, beliefs and conventions.

In the contemporary literature on social ontology, we have seen, there is a debate about the nature of the connection between background reasons and grounding. While some consensus exists on there being a difference between the ‘*move-making* facts’ grounding different social kinds on the basis of rules, on the one hand, and the ‘*rule-setting* facts’ supporting the existence of such rules, on the other, there is controversy over whether the type of metaphysical relationship is the same.¹²⁰

As we have mentioned, Brian Epstein claims that there is a peculiar kind of metaphysical link between rule-setting facts and the existence of social rules — a link, not corresponding to grounding, which he calls *anchoring*. (Mikkola 2017). Anchoring provides *glue* that holds a social kind together through the exposure of the facts that justify its grounding conditions (Epstein 2015, 58). According to Epstein, there are different theories about anchors, and most of them detect the anchoring facts in conventions, shared beliefs, and social practices.¹²¹ For instance, on this view, the anchor of a given institutional fact in John Searle’s theory is the

¹¹⁸ These rules and facts should be considered as complementary in constituting the social property of legal personality.

¹¹⁹ As noted for Epstein, these are not constitutive rules à la Searle, but frame principles that have different properties.

¹²⁰ The debate between Brian Epstein and Jonathan Schaffer revolves around this distinction (Schaffer 2019, 749; Epstein 2019, 768).

¹²¹ In particular, Epstein cites David Hume and John Searle’s theories. Hume believed that facts such as ‘being the owner of a piece of land as first occupant’ were anchored in shared beliefs and expected benefits regarding the rule of first occupancy. Searle, as is known, anchors institutional facts to collective attitudes of acceptance. See (Epstein 2014).

community's collective acceptance of the corresponding constitutive rules (Searle 1996). So, the 'anchors' ensure that there is a grounding connection between priority facts and nonpriority facts (e.g., institutional facts), according to the constitutive rules.

Jonathan Schaffer, on the other hand, insists that for both sets of facts — 'move-making' and 'rule-setting' facts — the underlying metaphysical relation is grounding, and that, together, the two sets of facts contribute to grounding a social kind. On this view, sometimes labelled 'conjunctivism,' social kinds are grounded simultaneously in the facts stipulated by a social rule (R) — e.g., a piece of paper issued by the ECB is a euro banknote — and in the facts that cause the social rule to obtain — e.g., collective acceptance of (R).

Although I find the arguments on the distinction between grounding and anchoring convincing, it is not essential to embrace one metaphysical model or the other on this specific point. I think it is fruitful to consider what they have in common, that is the distinction between two layers in the structure of social (and legal) kinds, i.e., the level of use conditions/implications of a kind and the level the corresponding background reasons. In the following analysis, and with the aim of providing a conceptual framework of personality, I shall refer to the first as the *institutional layer* and to the second and the *meta-institutional layer*.

As further explained in the following pages, on the one hand, the set of use conditions (a) and legal implications (b) provide the institutional layer, i.e., the move-making facts of personality,¹²² while the background reasons (c) make up the meta-institutional layer, i.e., the rule-setting facts. Background reasons provide answers to questions like the following: what makes it so that the fact of 'x being a company' *grounds* the fact of 'x being a person in law' or, otherwise stated, what grounds, e.g., the rule according to which corporations are persons in law?

Finally, even if personality — like other legal concepts — is aptly described by its 'use conditions' and 'legal implications,' to answer the question whether certain entities are suited to acquiring personality in law — whether they *should* be granted personality — we need to investigate the reasons underlying the ascription of personality. The longstanding theoretical debate on the nature of legal personhood is indeed mainly a

¹²² As will be made clear in the following sections, there are good reasons to believe that preconditions and consequences complement the institutional status.

debate about background reasons, i.e., the anchors (or second-level grounds): some authors believe them to be independent of the legal context, while others view them as mostly systemic and contingent.

4. Theories of Legal Personhood: Legalism vs Realism and Pluralism vs Monism

There are many approaches to legal personality. I will distinguish them here along two axes: one contrasting legalist and realist approaches, the other contrasting pluralist and monist approaches.

Legalists share the idea that legal personality is basically an artifice through which positive legal systems create a bearer of rights and obligations. They focus on positive law rules specifying the ‘use conditions’ and ‘legal implications’ of personality and on the specific function it plays in a legal ecosystem. As Lawson points out: “All that is necessary for the existence of a person is that the lawmaker, be he legislator, judge, or jurist, or even the public at large, should decide to treat it as a subject of rights or other legal relations” (Lawson 1957, 909).

As is known, this way of thinking was most clearly expressed in the theory of Hans Kelsen. According to Kelsen, an entity is a person under the law when this entity is an addressee of legal norms, i.e., when an action by that entity may trigger a sanction, or penalty, against it. An entity’s legal personality consists in the set of norms that apply to it. Hence, the core of legal personality lies in ‘imputation,’ as happens in the process consolidating legal effects under an autonomous centre of interest within a legal system. For Kelsen, all legal norms govern human conduct; when a collective entity is personified, the regulated actions of certain individuals (those acting as organs or agents of the collective entity) are ascribed to the collective entity (Kelsen 1945).

Another distinctly legalistic view is that of H. L. A. Hart, who focused on the linguistic use of legal concepts. Hart thought that nothing essentially corresponds to legal terms, since their core meaning has to be found in the function they perform in legal contexts, rather than in some metaphysical foundation: “legal words can only be elucidated by considering the conditions under which statements in which they have their characteristic use are true” (Hart 1984, 37). The meaning of ‘person’ in law is therefore the result of the linguistic practice that jurists are committed to.

I think that the legalistic view is meaningful as long as it refers to the grounding relation strictly understood, namely, to the use conditions (a) and legal implications (b) of legal personality. Such conditions and implications are indeed established by each legal system. However, an exclusive focus on the legalistic perspective may lead to disregarding the background conditions for ascribing personhood, and thus to an insufficient understanding of the link between personality, on the one hand, social attitudes, ethical values, political ideals, and, as I believe is also the case with artificial intelligence systems, socio-technological opportunities and risks. This limitation of the legalistic view is only partly overcome by broadening the legal perspective so as to include not only the grounding conditions explicitly fixed by legal rules, but also other sources of law, such as legal conventions, shared legal principles, and pragmatic-instrumental reasons pertaining to legally protected interests and values.

In the legalist approach, MacCormick's neo-institutionalist theory of personality proposed to integrate rationales of a non-strictly endogenous character: "The very business of imputing acts and rights or liabilities as consequences of acts to 'persons' is a business whose sense is that of motivating possible actors and of securing interests for potential sufferers from actions. That requires us to suppose that at least some actors could in fact be motivated to some actions, and at least some sufferers protected of their interests, by reason of the way people orient themselves to law by thinking of themselves as its persons, or as participants in its personified agencies" (MacCormick 1988, 373). By this MacCormick, in line with the legalist matrix, does not infer that to be a person is to be considered as a simple brute fact or that there is a criterion (or a set of criteria) to attribute this status in a manner unrelated to the socio-juristic practice of imputation of acts and interests.

Contrary to legalists, realists believe that essential (necessary and sufficient) conditions exist which justify legal personality and that these conditions have universal application, independently of the content of particular legal systems.

Since there are different views on the essential precondition for personality, as Ngaire Naffine points out, we have at least three subtypes of realism, each seizing on different properties (Naffine 2009, 20):

Religionist. Since all human beings are equally sacred, they are worthy of legal protection through the conferral of personality regardless of their cognitive abilities.

Naturalists. What matters for legal personality is that human beings have sentience and share a common destiny. This viewpoint can, of course, be used also to promote the personality of nonhuman animals possessing sentience; relevant properties from this perspective are: self-awareness, sensitivity, goal-oriented behaviour, capacity to suffer, sense of future and past, curiosity and so on.

Rationalists. Legal personality is suited to those who are cognitively able to exercise the corresponding rights in a rational manner. What matters is the ability to engage in legally relevant interactions and to sensibly assume responsibility for such acts. As Naffine correctly points out, these thinkers tend to restrict legal personhood to a narrow circle of individuals, thus excluding ‘marginal’ human beings (e.g., very small children or severely mentally impaired adults).¹²³

To these three categories illustrated by Naffine, I think we can add a fourth class of realists, not least because of its prevalence in the debate on the personality of natural entities:

Relationalists. The contours of legal personality are not determined by the possession of specific innate traits or capacities, but rather by membership in a ‘community of recognition’, where human entities form their identity and values in part by interacting with nonhuman entities (e.g., the environment and its constituent entities) (Silva 2017). Thus, for instance, communication skills, vulnerability or dignity can be seen as relational properties (Consiglio 2019).

It seems to me that this view is endorsed by Chopra and White when discussing personhood of artificial entities: “Personhood is a status marker of a class of agents we, as a species, are interested in and care about. Such recognition is a function of a rich enough social

¹²³ The philosophical roots of this school of thought trace back to Locke and Kant; many are the modern exponents of this orientation (Gray 1921; Moore 1984; J. Gardner 2003; Solaiman 2017).

organization that demands such discourse as a cohesive presence and something that enables us to make the most sense of our fellow beings. Beings that do not possess the capacities to enter into a sufficiently complex set of social relationships are unlikely to be viewed as moral or legal persons by us. Perhaps when the ascription of second-order intentionality becomes a preferred interpretational strategy in dealing with artificial agents, relationships will be more readily seen as forming between artificial agents and others, and legal personhood is more likely to be assigned [...] The question of legal personality suggests the candidate entity's presence in our networks of legal and social meanings has attained a level of significance that demands reclassification. An entity is a viable candidate for legal personality in this sense if it fits within our networks of social, political, and economic relations in such a way it can coherently be a subject of legal rulings" (Chopra and White 2011, 187).

Anyway, all these four interpretations share the idea that in order to identify the spectrum of legal subjects, one must look beyond the boundaries of positive law, because that is where the justification for this status is to be found. The realist perspective, regardless of the merit of particular theories falling under this umbrella, disregards the role of positive legal systems in shaping the conditions and implications of rationality. We can view it as expressing natural law perspectives, meant to override different choices made by specific legal systems. Or we may view it as advancing specific theories on the background reasons for ascribing rationality, theories in light of which to critically consider the choices made within specific legal systems.

Along the axis of pluralism vs. monism, we might distinguish approaches that accept that different conditions may exist for conferring rationality — with different implications and in light of different background reasons — from approaches on which, on the contrary, a single set of conditions always triggers the same implications, and on the same rationale. Obviously, intermediate positions may also exist.

Legalistic approaches are generally pluralistic, as they recognise whatever different grounds for personality may be recognised by different present, past, or even future legal systems.

Realist approaches, on the contrary, skew toward monism, usually focusing on a single ground or rationale for personhood. Probably an exception to this categorisation is posed by the class of relationalists, who do not appear to be monists given the set of systemic and

interactional factors they employ to delineate the contours of legal personality.

I believe that a pluralistic approach should be adopted, not only at the level of the grounding relation but also at the level of the background reasons (which anchor or ground in a structural fashion). In fact, different justifications may support the conferral of personality of different types of entities, and in different areas of the law. It follows that we have at our disposal an armoury of metaphysical anchors of legal personality (I will then attempt to represent them through meta-institutional legal concepts).

Thus, for instance, criminal law emphasizes cognitive attitudes and rational behaviour, while medical law appears to be influenced by religious or moral values connected with the sacredness of life (this is what emerges, for example, when we turn to issues like the end of life, abortion, and the legal subjectivity of the unborn) (Naffine 2009, 164). As a result, there may be good reasons to consider an entity a legal person under medical law, but not under criminal law. Finally, as previously emphasised with the ‘argument from marginal cases’, even the criterion of rationality does not provide a univocal guide for attributing moral or, in the case at hand, legal statuses.

Legal personality is therefore not a unitary concept: whether an entity is eligible for personhood and, if so, what kind of personhood it is eligible for, is a question that needs to be contextualised (Tur 1987). The need for contextualisation, however, does not debunk the idea that the ascription of such personality requires reasonable background reasons. Such reasons may be linked primarily to economic purposes, pertaining to the efficient management of resources and activities, as is the case for collective entities, but even more so for impersonal legal platforms like single-member or memberless companies.¹²⁴

5. Legal personality theories reviewed according to the anchoring-grounding diagram

Besides accurately representing the structure of the legal system, according to Epstein the GAF, by virtue of the grounding-anchoring diagram, can frame some differences between the main legal

¹²⁴ Here the distinction proposed by Visa Kurki between 'legal persons' and 'legal platforms' seems also relevant (Kurki 2019, 133).

philosophical theories. In particular, the debate on the nature of law can be conceived as a debate on the *grounds* for primary rules: e.g., are rules commands or are they reasons for action, and so on. The debate on sources or validity in law, and the role of morality in these aspects, can instead be conceived as a debate on the *anchors* of primary rules and secondary rules: e.g. for *exclusive* legal positivism both rules are anchored in strictly social facts, as it holds that neither the existence nor the content of legal norms need moral facts (the so-called social sources thesis applies); for *inclusive* legal positivism, on the other hand, this applies only to secondary rules, since the existence of a legal rule in a legal system always depends on a social fact (the so-called social fact thesis), whereas moral reasons may determine the judgement of validity of primary rules, which are consequently anchored to both social and moral facts. More generally, any theory of the criteria for the validity of legal rules can be seen as a theory of the facts that anchor the primary rules; and any theory of the sources of law as a theory of the facts that anchor the secondary rules.

This discussion of the main legal philosophical currents mirrored with regard to theories of legal personality. Specifically, as anticipated, it is a debate that takes place on the level of *anchors* – of primary rules and secondary rules – i.e. on what set up (legal) rules, hence the grounding conditions, of legal facts and kinds rather than on the grounds themselves.¹²⁵

Legalists, who are essentially legal positivists, tend to see law and legal kinds as the product of social practice alone. Yet, with regard to conceptions of personhood, we might use the diagnostic tool of anchoring to distinguish *radical* from *moderate* legalists. For radical legalists, who believe that 'anything goes' for the attribution of personhood as long as it is justified by legal purposes, both the anchors of primary and secondary rules are strictly systemic-social in nature (they are valid by virtue of a social pedigree). One such model of legal personality is that of Kelsen. For moderate legalists, on the other hand, although the sources of the criteria for the validity of the status of a person under the law, i.e. the anchors of the secondary rules, remain systemic-social, the very criteria of legal validity, i.e. the primary rules, are not free from moral or extra-legal evaluations (conveyed for example

¹²⁵ There may be debate about the level of the grounds, but I think they mostly are effects of different understandings of the anchors.

through judicial interpretation). An example of a moderately legalistic reading of personality is that of MacCormick.

Perhaps the distinction between radical and moderate positions applies to realists as well: the radical ones share the idea that both primary and secondary rules about personality are anchored in specific properties, natural or spiritual facts – e.g. rationalists, naturalists, religionists – while the moderate ones seem to be committed to such statements only in terms of the anchors of secondary rules – e.g. relationalists.¹²⁶

6. Conclusion

In this chapter I have engaged in a *real definition analysis* of legal personality, while also trying to include the main elements of the doctrinal debate. Although a preference for the neo-institutionalist, and therefore moderately legalist, theory has emerged several times, the aim has been to promote a metaphysical structure of legal personality, as ecumenical as possible with respect to specific theories of personality.

I believe that, as argued, the potential of the diagnostic tools of metaphysical grounding and anchoring is manifold in the legal field. One reason for separating the two metaphysical investigations, not considered so far, may be to diversify the activity of the legal official, legislator or judge, from that of the legal theorist. The first seeks to set or specify grounding conditions – e.g., what counts as a bribery offence – to establish if actual facts satisfy certain founding conditions – e.g., whether something, a given course of conduct, falls within a legal category – and to explore possible worlds in which actual legal facts are grounded in alternative facts – e.g., what happens if we apply an alternative taxation regime to a particular legal entity (for Epstein the so-called model building project) (Epstein 2015, 99). The legal theorist, on the other hand, aims to investigate the criteria that make a certain legal norm, or a system of legal norms, valid, and thus which facts (social or otherwise) determine those criteria. This includes speculations on the sources and nature of law, the connection between law and morality, normativity, and other typical topics of legal philosophy. It follows that

¹²⁶ It seems to me that a relationalist would have no great problem recognising that the primary rules are anchored in the specific way the legal system operates (although this would perhaps problematise their inclusion in the class of metaphysical realists).

the legal theorist is more interested, though not exclusively, in second level facts (for Epstein the so-called anchoring project), i.e. those which trigger the existence of certain rules and conditions rather than others.

In the next chapter, an attempt will be made to develop a conceptual framework that is equally neutral with respect to conceptions of personality, but with the intention of bringing elements of theoretical clarity to the debate. Also on the debate that invests the legal personality of artificial intelligence systems.

IV. Combining Layers for the Scope of AIs

1. Conceptual and metaphysical knowledge

The meta-institutional conceptual layer fills the space left uncovered by reductionist/inferentialist views on concepts: it provides the background reasons for the legal institutions-concepts. Meaning those reasons that make it possible for there to be certain use conditions and normative implications for a given legal category.

Schwyzler's intuition was to some extent reproduced, with significant entailments for social and legal ontology, by Andrei Marmor's thesis about the distinction between *deep* social conventions – which “[...]emerge as normative responses to basic social and psychological needs” and “[...] serve relatively basic functions in our social world” – and *surface* conventions – which “often get to be codified and thus replaced by institutional rules”; while “deep-conventions typically resist codification (of this kind)” (Marmor 2007, 594).

Marmor provides several examples of social practices – e.g., games, art, and language – to illustrate the difference between surface and deep conventions; including the game of chess, which is the quintessential case of institutional practice (not coincidentally also used by Searle and Schwyzler). To avoid being repetitive I will take the art case, for Marmor the best illustration of deep conventions: both early medieval Christian-European and Islamic art are linked to an underlying convention that artistic practice is a functional tool for glorifying God and its deeds. This common core, however, is expressed through two very different visual art forms, the Christian narrative and representational, the Islamic purely

ornamental. So, while each of them are well founded on their deep convention, they are implemented by different (superficial) practices and rules. The parallelism between the institutional layer and surface conventions and between the meta-institutional layer and deep conventions is clear.

In any case, we must reiterate that the discourse moves along two tracks: the conceptual and the metaphysical (/ontological).¹²⁷

Conceptual knowledge is crucial for organizing and reasoning practically on institutional practices, e.g., to understand how they provide reasons for action. To this scope, we have seen how institutional reality, with its facts and kinds, can serve as a semantic referent of legal concepts. Legal concepts organise relevant information about institutional reality. MacCormick's neo-institutional theory of legal concepts is an example.

Metaphysical knowledge is crucial for understanding the origins and structure of social institutions, of which legal kinds and facts are a subset. According to the reading I am endorsing, institutions are systems of rules in equilibrium, that shapes human practices and expectations. In turn, the way institutions are manifested and structured through these rules is shaped by a set of deeper practices and beliefs that are not fixed by legal rules. In short, these latter facts, which we call *meta-institutional*, somehow determine the nature of the former.

The metaphysical explanation can rely on different models. I have used the notions of grounding and anchoring and, more generally, to the model proposed by Brian Epstein.

But just as metaphysically we give priority levels of existence and explanation, so we can identify distinct conceptual layers: here, too, institutional concepts seem to *presuppose* meta-institutional ones. Thus, the concept/status of legal personality as we discussed it in regulatory terms in the first part of this thesis is an institutional concept, and can take different forms depending on the normative content set by the constitutive rules. Concepts (and values) such as intentionality, agency, self-awareness, curiosity, rationality, dignity, vulnerability, sentience or sacredness of life are meta-institutional, providing justifications and rationales that prove the dependence of legal practice on wider social practice.

¹²⁷ Since I have also discussed ontology in relation to compositional theories of meaning and concepts, this confusion should be avoided.

In this work, the focus is on legal concepts (and facts), but nothing excludes that such a distinction can also apply to other institutional practices such as, for example, economic, religious, recreational, cultural, political and so on.

So, even if there is not consensus in literature about the strong connection between conceptual and metaphysical dimensions, it seems to me that the way conceptual knowledge is framed may reflect some ontological features of institutional reality. In this sense, I would agree with Jackson's remarks about the epistemic value of conceptual analysis for metaphysical enquiry: "Although metaphysics is about what the world is like, the *questions* we ask when we do metaphysics are framed in a language, and thus we need to attend to what the users of the language mean by the words they employ to ask their questions. When bounty hunters go searching they are searching for a person and not a handbill. But they will not get very far if they fail to attend to the representational properties of the handbill on the wanted person. These properties give them their target, or, if you like, define the subject of their search. Likewise, metaphysicians will not get very far with questions like: Are there *Ks*? Are *Ks* nothing over and above *Js*? and, Is the *K* way the world is fully determined by the *J* way the world is? In the absence of some conception of what counts as a *K*, and what counts as a *J*" (Jackson 1998, 30).

2. Meta-institutional *anchors* institutional?

However, it is perhaps not yet sufficiently clear how the institutional and meta-institutional conceptual layers relate to each other, as well as the relationship in which they both stand with the constitutive rules, in the conceptual framework I am presenting.

Up to this point, there has generally been talk of some kind of priority relationship whereby the meta-institutional dimension precedes the institutional one; or that the latter is somehow dependent on the former. The relationship of priority, or fundamentality, can be expressed, as we have seen discussing about the structure of social reality, through the metaphysical notions of grounding or anchoring (the latter being more controversial).

I would like to address the question of whether these notions are applicable to the relationship between institutional and meta-

institutional layers; and then also whether they account for the conceptual dimension.

It seems reasonable to assert that an institutional fact such as 'Herbert has legal personality' (1.1) is partially *grounded* by the constitutive legal rule (R) jointly with the fact that satisfies the grounding conditions established by it, e.g. 'Herbert is a human being' (1.2). In fact, Herbert is a person of law (1.1) in virtue of (R) and (1.2), not on causal factors, but according to the metaphysical features of the constitutive and explicative grounding relation (e.g. synchronicity, asymmetry, irreflexivity, and so on).¹²⁸ Here we are not interested in discussing whether it is more correct to use the notion of constitutive rule or frame principle, nor whether the grounding relation is suitable to describe the metaphysical structure of legal facts or whether it is necessary to revise it in the light of a relation that better accounts for the normativity of these facts (one possibility is the so-called *normative grounding*, see Part II, Cha. II, sec. 5.3.).

But what is the metaphysical relationship between the institutional and meta-institutional layers? As mentioned, the relationship occurs between facts brought about by a set of rules – i.e., institutional – and facts, about practices and beliefs, that are not determined by these rules, but that rather put in place them to govern the former – i.e., meta-institutional. Put in this way, the relationship between the two layers seems similar to what Epstein regards as an anchoring relation. Yet, to test whether this hypothesis is correct, thus endorsing the metaphysical distinction between grounding and anchoring, we need to sift through the alternatives and look more closely at the peculiarities of anchoring.

The first alternative relation we can consider is *supervenience*, i.e. a logical and/or metaphysical relation of (necessary) covariance between sets of properties. It would seem to exclude the recurrence of supervenience between the two layers, neither in one sense nor the other, as Roversi correctly observes: “[...] the institutional layer does not supervene on the meta-institutional layer. Imagine chess being modified by its game designer, yielding Chess 2. Chess 2 could very well differ from Chess 1 in its rules but not in its assumptions about victory,

¹²⁸ The same relationship can probably also be expressed in terms of existential dependency, whereby X depends existentially on Y if (and only if) X exists only if Y exists. However, this is not a particularly explanatory relation.

cheating, sportsmanship — indeed, this is almost always the case when two games are compared” (Roversi 2021, 8).¹²⁹

Another metaphysical connection that can be considered among the candidates is that of *existential dependence*: X depends existentially on Y if, in order for X to exist, it is necessary that Y also exists (or “it is an essential property of X that it exists only if also Y does” (Fine 1995, 272). Although one can recognize some form of existential dependence of the institutional level on the meta-institutional level, the problem with this relation seems to be its explanatory weakness. Thus, for example, the fact that Joe Biden is President of the United States may depend existentially on the fact that $2 + 2 = 4$ (Fine 2012).

Finally, we should consider the grounding relationship, already discussed above. It can be argued that the institutional layer is grounded by the meta-institutional layer, as an institutional fact exists *because of* – or in *virtue of* – the meta-institutional one: the fact that X is a person of law would be grounded – perhaps *partially* – by the fact that, e.g., X has intentionality.

Yet, if we accept the metaphysical validity of the distinction between grounding and anchoring, the relationship between institutional and meta-institutional can be a dimension to which the same can be applied. We know that anchoring, as well as grounding, is the *metaphysical reason* for the anchored fact to obtain, and what is anchored is typically the constitutive rules (or frame principles): “[...] anchoring is not the metaphysical determination of a fact. Rather, it is the metaphysical determination of a way facts of a certain sort are grounded” (Epstein 2019, 239). Similarly, meta-institutional facts do not appear as metaphysical reasons directly of institutional facts, but of the set of rules that creates institutional facts.

One way in which Epstein argues for the different metaphysical quality of the anchoring relation from that of grounding, and which can be applied in this situation, is that social (institutional) facts and kinds can be *exported* independently of the relevant anchors, but not of their grounds. To export a social fact/kind means to test the recurrence of its grounding conditions in worlds that lack its anchoring facts. According to Epstein, this implies that it makes sense to ask whether, for example, Caligula satisfies the grounding conditions for being a *war criminal* even

¹²⁹ The same reverse, i.e. meta-institutional, does not occur institutionally unless there is a global change in institutional practices.

in the absence of the facts that anchor the social kind (e.g., the Geneva Convention, the International Criminal Court or, axiologically, democratic accountability). While the same shift would not seem to make as much sense if, as conjunctivism proposes, one equates grounds and anchors, as social facts and kinds, e.g., war criminal, would only be obtainable in the presence of both of the aforementioned metaphysical roots.

I believe it is appropriate to consider the relevance of this line of reasoning to the distinction between institutional and meta-institutional levels as well. In a sense, the same thought experiment that Schwyzer formulates to flesh out meta concepts demonstrates that an institutional practice such as the game of chess, made up of certain fixed constitutive rules (which set grounding conditions), can be exported to contexts where the relevant anchors do not recur, the society of Ruritarians. The same can be argued for legal categories: we can sensibly ask whether a slave or a woman during the Roman empire, animals or environmental entities, unborn children, and artificial intelligence systems are persons in law even where the anchors – here sometimes called background reasons – that put in place the constitutive rules of personhood status are missing. What matters is rather that the anchored conditions are met.

In other words, in the absence of the relevant anchors, like the fact that the civil law code has been promulgated – or of the Universal Declaration of Human Rights – or of the social recognition of equal dignity for men and women, we can still ask ourselves whether the grounding conditions for legal personality have been met. None of the anchoring conditions fall within our conception of the institutional dimension since they are not fixed by constitutive rules, but it is precisely the facts (or kinds) that put the rules in place.¹³⁰

The analogism between meta-institutional dimension and anchors probably does not apply to everything Epstein understands by anchors. In fact, precisely by applying his model to the case study of law, he distinguishes between two levels of anchoring. The first level – e.g.,

¹³⁰ This profile is actually more problematic: according to Epstein it is possible that the anchoring conditions are set by rules; instead of constitutive rules he considers them, as anticipated, frame principles. In any case, beyond the debate about the validity of frame principles, even accepting the thesis that anchoring conditions are fixed by some kind of social rule, we can agree that these rules are qualitatively different from those that fix the grounding conditions, the legal and therefore institutional facts (and kinds). They are at most Hartian rules of recognition.

judicial decisions, jury instructions, the fact that the legislature enacted the statutes, etc. – is made up of facts that tend to be *procedural*, whose place in the dimension of the meta-institutional is more doubtful. In a sense, deep anchors are more reasonably meta-institutional.

Although it certainly requires more exploration, this perspective on the metaphysical structure of legal facts and kinds seems well-founded to me.

3. Theoretical benefits of a multilayered ontology and intermediate legal concepts

In the previous sections, the inferential reading of institutional concepts was found to be consistent with an ontological apparatus that also includes the backstage of meta-institutional concepts. Now, what advantages derive from this joint framework? In pointing these out, it will be useful to distinguish the purely theoretical implications from the practical ones (especially from the focus on the legal status of AI).

The theoretical advantages — i.e., for a general understanding of legal concepts — are multiple. In the first place, the meta-institutional level shows that an institution's constitutive rules do not emerge out of nothing but take shape within a conceptual and 'semantical atmosphere' (Lorini 2018, 13). The teleological and axiological landscape influences institutional practices, e.g., concepts of 'moral agency' or 'moral patience' imposes some constraints on legal personality itself. It is thus possible to highlight some of the beliefs the legislator is committed to when making decisions about legal policy with respect to who should be recognised as having subjective legal positions and how. This point reinforces MacCormick's claim about the inadequacy of an account of institutional practices solely based on constitutive rules: "[...] unless you know the underlying principle or final cause of a given institution, it profits you nothing to know how an instance of it can be established" (MacCormick 1998, 319).

Second, a multilayered ontology shows that legal institutions are not self-contained, but rather that some external concepts are presupposed. An awareness of these interlinkages can play a significant role in legal cognition itself (Sartor 2009). This reinforces the idea that legal forms tend to be constantly adjusted to social practices (Deakin 2015). In addition, this multilayered ontology provides conceptual gateways between the different degrees of existence of institutions in

MacCormick's theory (namely, between the social dimension and the juridical one).

Third, a multilayered ontology enables us to understand the merit of having general concepts, such as legal personality, covering, or potentially covering, disparate situations (e.g., humans, children, companies, animals, AI systems). In fact, the meta-institutional level brings into account interests and values —which evolve depending on the period and the social system in question — grounding distant circumstances for the possible ascription of personality. Occasionally, this background may highlight affinities or distinctions, enabling or disabling analogical connections. As a consequence, the coexistence of *thick* and *thin* notions of legal personality also becomes comprehensible.

Fourth, a theory of legal personality should address not only the 'use conditions' and 'normative implications' of such a legal status (see *supra*, sec. 3), but also its 'background reasons', that is to say, what facts bring into play the rules framing the conditions of legal personhood. We cannot here discuss pluralism or individualism about grounds or anchors;¹³¹ what matters for us is that meta-institutional concepts incorporate background reasons highlighting the deeper facts and values that are presupposed by legal institutions and their constitutive rules; that is to say, they describe so-called 'rule-setting facts'. On this model, the metaphysical structure of legal personality bears at least two different relationships: an *anchoring* relationship – or *structuring* grounding (if one takes Schaffer's position as valid) – between facts at the meta-institutional level and the constitutive rules of legal personality; and a *triggering* grounding between move-making facts and the institutional fact/property of legal personality (Schaffer 2019).

Detecting anchors (or structuring grounds) is useful for a better justification of rules in legal enforcement. They point out prior general values and social conventions able to shape our choices in the legal sphere. Even though legal concepts are constituted by inferential links, if we are to justify particular inferences and then commit to their applicability, we need these concepts and inferences to be consistent with some deeper reasons. Law is able to provide reasons for action and motivate individuals thanks also to the involvement of deeper

¹³¹ For an interesting reading of pluralism and individualism about grounds and anchors —in partial disagreement with Epstein— see (Guala 2016).

conventions rooted in the social practice in which the law is embedded (Marmor 2007).

Finally, closely related to the previous point, meta-institutional concepts can be employed in the legal field to *test* and *reformulate* the content of institutional concepts (Roversi 2014). In the case of personality, such concepts can help us question who or what ought to have the legal protections and empowerments linked to personality and in which way.

4. On the existence of an intermediate (or quasi-institutional) layer: legal subjectivity

This last point bears some further elaboration. Sometimes the abovementioned review is carried out indirectly: rather than referring explicitly to meta-institutional notions, auxiliary concepts are used to bridge the indeterminacy of meta-level pro and con reasons with the (relative) determinacy of legal rules. The function of these auxiliary concepts is similar to that which in moral philosophy is served by what are referred to as *midlevel principles*, which help to justify rules and particular judgments by specifying or balancing higher-level principles (Bayles 1986). By going back and forth between deep reasons and the particular assessments, we make in concrete cases is what characterizes the method of *reflective equilibrium* (Rawls 1971; Daniels 1996). Through this method, coherence is sought ‘among the widest set of moral and nonmoral beliefs by revising and refining them at all levels’ (Daniels 1996). As we shall see in the next section, auxiliary terms can also be used to achieve a kind of equilibrium in the legislative debate.

In the area of legal personhood, *legal subjectivity* may be viewed as an auxiliary transitional concept of this kind, partly overlapping with the thinnest notion of personality (having, or having the ability to acquire, some legal rights or duties); I will come back to this in a moment. This concept is meant to capture, at least in Italian legal commentary and jurisprudential debate, potential eligibility for personality status. Legal subjectivity is thus mostly used to underpin some entities’ merits or competence in relation to legal rights, frequently in a mitigated or conditioned manner (consider, for instance, condominiums or unborn children). However, legal subjectivity is not an institutional concept proper, since it does not refer to a precise package of legal positions.

Nor can it be said to be a meta-institutional concept, as it depends on the features of legal practice, including as set out in its constitutive rules, rather than shaping them in a particular way.

With the notion of legal subjectivity, we can proceed on either a descriptive stance or a normative/programmatic one: in the former case we point to the minimal personality status, i.e., we point out that some entities are already recipients of at least one legal position (like a *de facto* entitlement); in the latter case we claim that these entities should acquire some broader personality status.

The former case, when entitlement to rights and duties arises even though there is no formal recognition of the status of a person (*de facto*), is often considered to characterise the legal standing of animals in many legal systems. Sunstein argues that animals do in fact possess a sphere of legal recognition of the personality type, in the form of a presumed guarantee or immunity, and that the problem is not one of legitimising this standing but rather of enforcing it. On this point, Naffine also: “It is that animals may *already possess* some of the rights which we associate with persons even though we do not tend to think of them as such. Accordingly, the property status of animals is not absolute and certainly it does not, conceptually, have to be absolute. Sunstein therefore questions the accuracy of the stark dichotomy between human persons and animal property depicted by the animal advocates. These two founding legal concepts, persons and property, he suggests, do not operate in this rigid inflexible manner.” (Naffine 2009, 138).

I have previously argued that the intermediate – or *quasi-institutional* – notion of legal subjectivity may partially overlap with the thinnest version of legal personality. Here, however, I employ a use of the distinction between *thin* and *thick* concepts, deferential to its proper and technical use in metaethics; rather than the legal-descriptive use I made earlier. In metaethics, this distinction concerns the content of normative concepts, where a thin term such as 'right', 'wrong' or 'good' is suitable for describing a wide range of conduct and situations, in contrast to thick terms such as 'generous', 'courageous' or 'cruel' which appear more descriptively specific and loaded with meaning. The two classes of terms imply then different descriptive commitments (both evaluative and non-evaluative) for the speaker.

Similarly, we can consider mere aptitude for entitlement to subjective legal positions – i.e., here legal subjectivity – as a term of *thin* prescriptive and descriptive content. Indeed, the aptitude to which legal subjectivity

refers does not commit to any particular conformation or intensity – hence to any particular package of rights, duties or other legal positions – of the status of a person of law.¹³² In its programmatic version, legal subjectivity calls for ensuring some form of legal recognition or protection for a given entity, leaving the legislator or policy-maker free to carve out the status as he or she sees fit. Is a notion through which a preliminary legal relevance is assigned to interests, capacities and values – even if only instrumental to human interests – not yet formalised through the creation of a centre of imputation of legal positions. Legal subjectivity remains a tool for conveying the specific intentions of the legal system with respect to factual situations, and to which the recognition of meta-institutional practices and values is not sufficient.

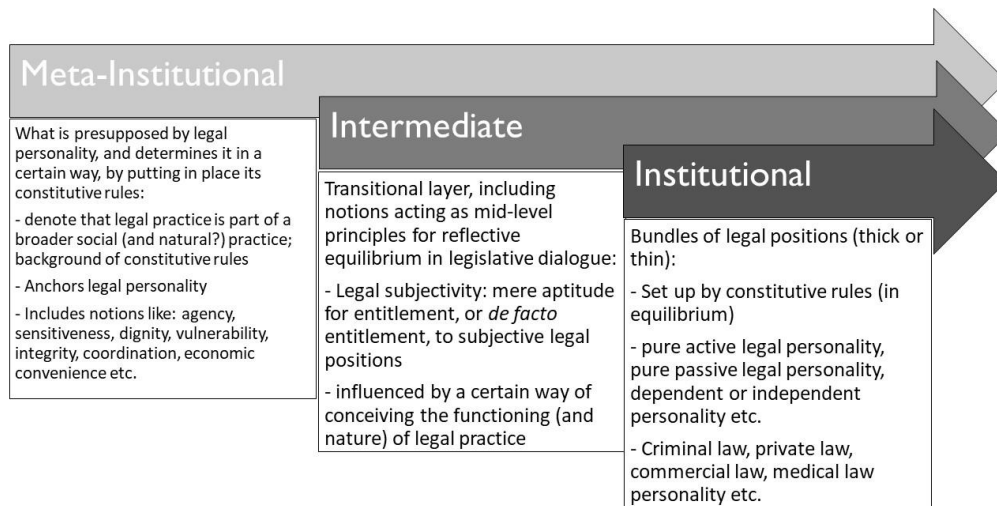
From this perspective, the institutional concept of legal personality is instead a *thick* term. Even if the content of legal personhood can be more or less articulated (more or less thick, as previously intended), the expression refers to the entitlement to definite subjective juridical positions. The legal designation is official and the legal system will not need to seek further qualifications for entities that enjoy it (Canale 2015, 33).

As with the meta-institutional level, the type of metaphysical relationship between the intermediate (or quasi-institutional) level and the institutional level deserves further investigation. One might think that in this case too we find ourselves in a kind of anchoring relationship – according to the diagram proposed by Epstein – with the difference that here anchoring is of a more procedural character, or of a second level. It would therefore be part of those anchors among which Epstein includes judicial decisions and trial results, as legal subjectivity describes the disposition to hold legal positions according to the features of a given legal practice, in line with the modalities of the legal system. But this profile calls for further study.

This is, more or less, the picture of the neo-institutional ontology of (the concept of) legal personality – and of other legal concepts:

Figure 6. The neo-institutional layers of legal personality

¹³² On the applicability of the distinction between thick and thin terms to legal personality (Canale 2015).



Yet, even if untangling these two terms is theoretically sensible, is there a practical reason to associate a concept with the mere potentiality of possessing juristic positions? I will try to argue this in the next section.

5. Reflective equilibrium between institutional layers: with a little help from mid-level principles

A way of justifying legal policy choices that balances the set of meta-institutional and institutional layers can be inspired by the model of reflective equilibrium (Goodman 1955; Rawls 1971; Daniels 1996). This point requires a brief introduction to the method of reflective equilibrium and the role mid-level principles play in achieving it.

The first reference to reflective equilibrium can be found in the philosophy of logic, specifically in Nelson Goodman's effort to justify the conclusions of deductive (or inductive) inferences in the light of general rules: "The point is that rules and particular inferences alike are justified by being brought into agreement with each other. A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between

rules and accepted inferences; and in the agreement achieved lies the only justification needed for either” (Goodman 1955, 64).

Later the same approach was adopted by John Rawls in *A Theory of Justice*, to understand how we draw practical decisions from general moral theories (Rawls 1971). Rawls claimed that the way in which we consistently reach moral decisions and evaluations results from testing considered judgments against (sets of) more general principles. If this mutual adjustment is successful, i.e., if specific judgments are aligned with general moral principles and theories (and vice versa), then we will achieve a reflective equilibrium among our moral beliefs; if, on the other hand, we fail to achieve such consistency, perhaps because we are unwilling to accept the practical or theoretical implications of a certain principle, then we will be led to revise the appealed principles (or judgements).

Rawls also distinguishes between two forms of reflective equilibrium, a *narrow* and a *wide* one: while in the former case the process of justification takes place between a set of considered judgments and moral principles belonging to a fixed theoretical background, e.g. a single conception of justice, in the latter, justification is global, in the sense that the considered judgments are to be tested against moral principles belonging to variable and competing theories on, e.g., the conception of justice (Rawls 1974).

A further development of the reflective equilibrium, not limited to moral beliefs, was carried out by Norman Daniels as a method for justification in ethics (Daniels 1996). According to Daniels, reflective equilibrium is an attempt to find consistency among three sets of beliefs:

- (a) a set of considered moral judgements
- (b) a set of moral principles
- (c) a set of relevant background theories (Daniels 1996, 22)

For Daniels, this theoretical background, which characterizes wide reflective equilibrium, can consist of both moral and non-moral beliefs. Among the moral ones can be included normative conceptions of equality or human nature, particularly relevant to our topic; among the non-moral ones Daniels includes both theories imported from other social sciences of a not specifically normative character, e.g., general social theory, and epistemological theories, e.g., a certain theory of mind. To avoid committing a kind of naturalist fallacy, it must be emphasized

that the task of non-moral theories is not to validate or revise moral principles, but to probe the logical or empirical feasibility of certain moral principles.

A decisive contribution to the adjustment process described in these various versions of the reflective equilibrium is made by the so-called mid-level principles. Mid-level principles hold an intermediate position between (fundamental) principles and considered judgements, presenting a lesser degree of abstraction than the former but undoubtedly greater than the latter: “Mid-level principles are subordinate to fundamental norms, and their justification usually refers, directly or indirectly, to a fundamental norm” (Bayles 1986, 50). Thus given the fundamental principle, e.g., of freedom, a medium-level principle is that ‘freedom must be respected’.

Mid-level principles can both derive from the mediation between fundamental principles and specific rules, and from a balancing process that is entirely internal to higher-level principles. In the second situation one can speak of *comparative* mid-level principles: “An example is the principle of clear and present danger, according to which a government may not regulate speech unless it can show a clear and present danger. This principle is based on balancing the higher-level principle of free speech against other considerations, and thus counts as a comparative mid-level principle” (Moreso and Valentini 2021, 575). Another interesting classification of mid-level principles is that between *substantive* and *procedural*: while the former provide substantive solutions – what deserves legal protection (e.g., principle of freedom of political opinion) – the latter provide instrumental solutions - how to provide protection (e.g., principle of proportionality).

Against this background, the functional role of mid-level principles, as widely discussed in both moral and legal philosophy (Bayles 1986; Henley 1993; Sunstein 1996; Audi 2004; Blankfein-Tabachnick 2013), is to solve various problems of justification and coordination between moral (and non-moral) beliefs: they provide an intermediate ground for agreement that, by not committing to fundamental principles, offer a strategy for converging on values with some degree of generality rather than on mere specific judgments or rules. In this sense, mid-level principles are auxiliary tools for reflective equilibrium.

5.1. Back and forth between neo-institutional layers

In the (neo)institutional ontology of legal concepts provided in this thesis we find some of the instruments described so far. In particular, we can see the meta-institutional level as the most abstract, equivalent to the level of ultimate principles, and the institutional level as the most specific, equivalent to the level of individual rules, cases and judgements. In fact, we maintain that at the meta-level, certain concepts such as dignity, freedom, rationality, agency, etc., (partly) determine the institutional status of legal personality; in other words, the way in which it is framed through packages of legal positions.

However, it has also been emphasized that we do not move directly from meta-concepts to institutional ones, and that the justification of the latter also follows a transitional phase in which the generality of the former adjusts to the practical needs of the latter. We have called this institutional level as intermediate, and visualizing with the notion of legal subjectivity the balancing that occurs between beliefs that abstractly justify (or would justify) the conferral of legal personality on certain entities and the peculiarities of legal practice.

In this sense, legal subjectivity plays a role analogous to that played by mid-level principles. As conceived here, legal subjectivity has the characteristics of a *derivative*, rather than comparative, and *substantive*, rather than procedural, mid-level principle; but this last point can be debated, it is not uncontroversial that it can perform sometimes substantive sometimes procedural functions.

Thus, we can represent the interaction between (neo)institutional levels of legal personality in this way:

- (a) *Meta-institutional* = Background beliefs of legal personality ¹³³
- (b) *Intermediate* = Fitness to hold legal positions
- (c) *Institutional* = Specific rules about legal personality

Again, mid-level axioms-here legal subjectivity-can be used not only to find coherence within an individual view or internally within a given background theory (*narrow reflective equilibrium*), but to test particular judgments against alternative and competing views afferent to the background (*wide reflective equilibrium*), i.e., about which extra-legal values anchor legal personhood. Whether it is a narrow or wide reflective

¹³³ These beliefs can be combined into theories.

equilibrium also depends on the reference context: e.g. whether it is an equilibrium set between different legal systems, perhaps characterised by different meta-values included of legal personality, or within the same legal system with equal meta values. Of course, there can also be intra-system wide reflective equilibrium, which is the case when the balancing takes place between different meta-values anchoring different legal personality statuses, e.g. those pertaining to criminal law and medical law personality, within the same legal system.

I believe that the strategic value of intermediate notions lies precisely in being able to mediate between different premises about personhood, therefore with divergences with respect to the meta concepts to which we are committed, in order to find an agreement also for the purpose of making decisions of legal policy such as extending (or not extending) the contours of legal personality. Mid-level principles can thus help find what Cass Sunstein calls *incompletely theorized agreements*, i.e. agreements that sometimes occur only on particular cases in the context of strong disagreements on abstractions, and sometimes occur only on an abstract level in the context of strong disagreements on particular cases. “So” as Moreso and Valentini point out “they sometimes also involve mid-level principles, despite disagreement about both general theory and particular cases” (Moreso and Valentini 2021, 569).

Ultimately, the type of consensus, or wide reflective equilibrium, that can be reached by means of the notion of legal subjectivity is of this kind: legislators converge on a thin notion, agreeing on a mid-level standard that allows them to start from different premises (e.g., what anchors legal personality) and/or pursue divergent regulatory choices (e.g., what legal personality consists of).

In the present case, given the coordination problem that characterizes the choice of law policy on whether or not to grant personality to AIs, a *modest consensus* can be built around the recognition of the aptitude for the entitlement of legal positions to AIs. More on this in the next section.

6. Applying a meta-institutional and an institutional perspective to AIs

The approach introduced, which distinguishes meta-institutional, intermediate and institutional layers, can be useful for examining the

question of the legal personhood ascribable to AIs, since it enables us to distinguish criteria of ascription at different levels in different contexts.

A meta-institutional perspective enables us to ‘test’ whether an entity possesses the general properties and attitudes that may justify an ascription of personality to it: cognitive capacity, self-awareness, vulnerability, dignity, the ability to have interests, moral patience, sentience, social role, the ability to communicate and cooperate, economic expediency, etc. Some kind of cognitive capacity, for example, may be one element supporting an ascription of personality to AIs, and another supporting element could lie in their ability to facilitate certain social or economic interactions.

However, none of the properties identified at the meta-institutional level is sufficient, separately considered, to ground legal personality, i.e., to determine whether such entities should be granted legal personality. To this end, we need to balance the advantages and disadvantages that would obtain if legal personhood — as a general ability to have rights and duties — were to be conferred on such entities. Consider, for instance, the multiple implications involved in abortion, inheritance, medical liability, etc., that could result once legal personhood is conferred on the unborn child. This judgement may be facilitated if we preliminarily test the candidates for legal personality by resorting to intermediate concepts, such as the idea of legal subjectivity. This midlevel review may also consist of a judgment of expediency, which may also eventuate in the claim to personhood being rejected. It may be concluded that certain entities we have classified as legal subjects may not require personality, since the protections, guarantees, or enabling conditions that are suited to such entities (e.g., unborn foetuses, animals, ecosystems, technological systems) may already be secured under different legal regimes that are better suited to such entities. For instance, it may be argued that sentient beings like nonhuman animals should not be qualified as legal persons, as they lack the cognitive capacity to understand their legal positions and act accordingly. On the other hand, it may be argued that cognitively capable beings, such as advanced AI systems, do not deserve the protection that is granted to legal persons, since they lack sentience, being unable to experience proper feelings. Or, from the same midlevel review itself, a judgement of expediency may point in the opposite direction, suggesting, for example, that there may be rhetorical and normative value in conferring

legal personality even where some kind of legal protection are already in place.

On the contrary, let us assume that we conclude that a category of entities should be granted legal personality, with an accompanying set of rights and duties (or with an opportunity to acquire them given the appropriate operative facts). In this case, we need to determine how we should configure legal personality of such entities — with what restrictions or extensions relative to the default idea of personality as the general ability to have rights and duties the patrimonial domain — in keeping with an adequate balance and adjustment of presupposed meta-institutional values. On this basis, an argument can be made to the effect that such entities can already be viewed as legal persons based on existing law (*de lege lata*) or that a change in the law is needed for personality to be conferred (*de lege ferenda*).

With regard to AI systems, the second approach (a change in the law) seems more plausible. It seems inappropriate to grant general legal personality to AI systems merely through legal interpretation, given the novelty of such entities and how different they are from those that have so far been granted legal personality (which differences make analogies highly questionable), and the important political implications of choices about the role that AI systems should play in society.

As long as legislators have not made that choice, the law may account of the existence of autonomous AI systems by relying on transitional concepts like that of legal subjectivity, which conveys the idea that certain entities, given their special features requires some legal recognition of their interests or capacities. What is more, the recognition of the legal subjectivity of autonomous AI systems may be used to build consensus – a kind of incomplete agreement as previously stated – around the need for certain kinds of AI systems to have a legal regime that takes their cognitive capacities, social functions, and accountability gaps into account, granting them some legal protections, or enabling them to autonomously and actively engage in certain legal activities, or just limiting stakeholders' liabilities. Such a regime could be introduced even as uncertainty or even denial persists concerning the attribution of personality.

The case of the unborn child has indeed been addressed in a similar way in Italy. There is no rule in the Italian Civil Code granting legal personality to the unborn, and a permissive abortion law is in force. However, by referring to certain rules which protect human life, in the

Italian Constitution and international law, the Italian judiciary has argued that damages are owed for injury to the unborn child. According to the Court, while the unborn child is not a (natural) person, he or she is a legal subject whose interests are to some extent taken into account by the law. The same holds for unincorporated partnerships or associations, which in several legal systems are accorded different levels of legal protection and different legal powers according to their purpose and composition. Interestingly, the same kinds of partnership or association may be conferred legal personality in one legal system but not in another.

As already mentioned the decision to grant personhood to AI systems can be affected by coordination issues (Part One, Chapter III, section 6). It is difficult to imagine that a single criterion could ever be used in Europe to determine whether an AI system should count as a legal person.

Indeed, for one thing, the technological heterogeneity among AI systems and the context of their use makes it impossible to resort to a single criterion to determine in what cases such entities could have personality. The conditions that make legal personality appropriate in one context (e.g., e-commerce) may be very different from those that make it useful in another (e.g., robots used in health care or in manufacturing). And, for another thing, we need to consider that in Europe different conceptions of legal personhood exist, both within the same legal system and between different legal systems.

This does not exclude, however, the possibility of the European legislator recognising a particular legal status (or different such statuses) for AI systems which satisfy certain conditions: the fulfilment of technical standards — e.g., transparency, fairness, explicability, safety, reliability — as well as certain levels of performance and autonomy, as defined by technical capacities and users' choices. It is desirable for the legislator to take a multifactorial approach in which multiple sociotechnical aspects combine to support the recognition of a special legal status for advanced AI systems, albeit in a gradual manner and without necessarily triggering full legal personality. In this sense, the risk-based approach that characterizes the AI Act currently under discussion in Europe may prove functional in classifying AI systems according to their socio-technical impact, and thus their legal implications; possibly also in terms of personality.

Such a status may come into shape when the users and owners of certain AI systems are partly shielded from liability – through liability

caps, for instance – and when the contractual activities undertaken by AI systems are recognised as having legal effect – though such effects may ultimately concern the legal rights and duties of owners/users – making it possible to view these systems as quasi-holders of corresponding legal positions. The fact that certain AI systems are recognised by the law as *loci* of interests and activities may support arguments to the effect that —through analogy or legislative reform— other AI entities should (or should not) be viewed in the same way.

Should it be the case that, given certain conditions (such as compliance with contractual terms and no fraud), the liability of users and owners — both for harm caused by AI systems of a certain kind and for contractual obligations incurred through the use of such systems — is limited to the resources they have committed to the AI systems at issue, we might conclude that the transition from legal subjectivity to full legal personality is being accomplished.

The combination of the meta-institutional and institutional perspectives, incorporating social and technological aspects, and supported by a flexible legal subjectivity, can mitigate the problem of coordination we took as a premise:

(a1) the meta-institutional level identifies the area of potential ‘electronic persons’, and may be instantiated as a sort of testing procedure;

(a2) the intermediate concept of legal subjectivity would not commit national lawmakers to a general conception or specific arrangements of personality (*reflective equilibrium*), and they would remain free to carve out the institutional status in detail in accordance with their legal systems (and values);

(a3) finally, this would also give legislators, national and supranational alike, the ability to determine whether a personality status for AIs will exist under civil, contract, or criminal law, or a combination of the three.

To conclude, the analysis of the background of institutional legal kinds and concepts makes it possible to discover theoretical tools that can be effectively implemented, as well as to support the conferral of thick legal status, such as general personality. Auxiliary concepts like legal subjectivity may help us to locate and specify the aspects that are

relevant to the legal status under consideration, while also making our set of moral intuitions, social beliefs, and conventions more stable for the purpose of their legal use. The mutual alignment between specific institutional rules and broader meta-institutional values contributes to the formation of a *wide* reflective equilibrium among divergent conceptions of subjecthood and personality across different legal systems as well as across different areas within each legal system.

CONCLUSIONS AND LIMITATIONS

AIs make decisions and adapt to the environment with increasing levels of autonomy from their human creators and users. This is a great opportunity, but it is also problematic when things do not go as they should. These issues specifically come into play when some of the conditions for attributing liability for the behaviour of AI systems may no longer be met by the individuals who design and/or use those systems. Often there are no short-term technical solutions to accountability gaps, but there are good reasons to bear the (social) risk of AI systems making unexpected and untraceable mistakes.

In this context, I tried to show the weaknesses of conventional civil law liability schemes (mainly tortious one) to deal with the consequences of failures due to the proper way AIs function. The legal remedy considered was the conferral of legal personhood, justified primarily by the need to fill the resulting liability gap in the most efficient way. The various forms that legal personhood status can take have been examined, as was the type of financial/asset structure on which the new, hypothetical, legal entity should be based.

In order to address some of the theoretical obstacles to realising the legal policy choice of conferring personhood on AIs, I then approached the question from a metaphysical and conceptual perspective. I have tried to answer the following questions: *‘what is it for something to be a person of law?’* and *‘what do we mean by the concept of legal personality?’*"

For both questions, I have tried to draw on traditional theories by updating them to contemporary positions, proposing solutions that help

us find a metaphysical and conceptual framework, as well as theoretical compasses, for using the notion of legal personality and finding political balances.

However, a number of issues remain open and cannot be effectively addressed in this thesis. Some of these issues are closely linked to the technological progress of AI systems and their actual use. Indeed, as the recent proposal for a legal framework by the European Commission suggests, the use of high-risk AIs could be severely limited or discouraged.

Also, technical issues remain open on how to provide an AIs with an asset. For instance, since the AIs can be sold to different users, one might ask whether it is preferable to equip the overall AI system with a single asset or to equip the individual services provided by the same system with as many assets. The two solutions require different economic assessments, with the latter one likely to be very costly.

Finally, other questions that remain unanswered have to do instead with the desirability of artificial agents with autonomous assets and independent capacity to act. In fact, it may be argued that agents capable of producing legally relevant effects in an unsupervised manner could prove, in the best-case scenario, unnecessary, in the worst-case, dangerous. In these circumstances, rather than favouring legal personhood, an insurance mechanism (perhaps compulsory) might be preferred, since it has similarities in the distribution of risks and costs and does not give excessive power to the AIs. This distortion can probably be governed by constraining the active positions of legal personhood, but this hypothesis has not been sufficiently explored in this thesis.

REFERENCES

- Acemoglu, Daron. «Why not a political Coase theorem? Social conflict, commitment and politics.» *Journal of Comparative Economics* 31 (2003): 620–652.
- Alpaydin, Ethem. *Introduction to Machine Learning*. Third Edition. MIT Press, 2014.
- Alschuler, A. W. «The Changing Purposes of Criminal Punishment: A Retrospective on the Past Century and Some Thought about the Next.» *The University of Chicago Law Review* 70, n. 1 (2003): 1–22.
- Ananny, Mike, e Kate Crawford. «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.» *New Media & Society* 20 (2018): 973 - 989.
- Anderson, Michael, e Susan L. Anderson. «Machine ethics: Creating an ethical intelligent agent.» *AI Magazine* 28 , n. 4 (2007): 15–26.
- Andrade, F, P Novais, J Machado, e J Neves. « Intelligent contracting: software agents, corporate bodies and virtual organisations.» In *Establishing the foundations of collaborative networks*, di L. Afsarmanesh, H. Novais, P. Analide, C. (eds) Camarinha Matos, 217-224. Springer, 2007.
- Andreotta, Adam J. «The hard problem of AI rights.» *AI & Society* 36 (2021): 19–32.
- Anscombe, Gertrude Elizabeth Margaret. «Modern Moral Philosophy.» *Philosophy* 33 (1958).
- Anscombe, Gertrude Elizabeth Margaret. «On Brute Facts.» *Analysis* 18, n. 3 (1958): 69–72.
- Aoki, M. *Toward a Comparative Institutional Analysis*. MIT Press, 2001.
- Argyrou, Aikaterini, e Harry Hummels. «Legal personality and economic livelihood of the Whanganui River: a call for community entrepreneurship.» *Water International*, 2019: 752-768.
- Armour, John, e Horst Eidenmüller. «Self-driving Corporations?» *Harvard Business Law Review* 10, n. 1 (2020): 96–97.
- Armstrong, D.M. *Belief, Truth, and Knowledge*. Cambridge University Press,, 1973.
- Arrieta, A. B., et al. «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.» *Information Fusion* 58 (2020): 82-115.

- Atkinson, Robert. «The Case Against Taxing Robots.» *Information Technology and Innovation Foundation*, 2019: 1-30.
- Audi, Paul. «Grounding: Toward A Theory Of The ‘In-Virtue-Of Relation.» *The Journal of Philosophy* 109, n. 12 (2012): 685–711.
- Audi, Robert. *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton University Press, 2004.
- Ayer, Alfred. *The Problem of Knowledge*. Macmillan , 1956.
- Azhikodan, A.R., A.G.K. Bhat, e M. V. Jadhav. «Stock Trading Bot Using Deep Reinforcement Learning.» In *Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems*, di H. Saini, R. Sayal, A. Govardhan e R. (eds) Buyya. Springer, Singapore, 2019.
- Banaś, Paweł. «Why cannot anything be a legal person? A critique of Kurki’s Theory of Legal Personhood.» *Revus, Symposium on the theory of legal personhood* 44 (2021).
- Barfield, Woodrow, e Ugo Pagallo. *Advanced Introduction to Law and Artificial Intelligence*. Elgar Advanced Introductions series, 2020.
- Bayern, Shawn J. «Are Autonomous Entities Possible?» *Northwestern University Law Review* 114 (2019).
- Bayern, Shawn, Thomas Burri, Thomas Grant, Daniel Häusermann, Florian Möslin, e Richard Williams. «Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators.» *Hastings Science and Technology Law Journal* 2 (2017): 135-162.
- Bayles, Michael D. «Mid-level Principles and Justification.» In *Nomos XXVIII: Justification*, di J. Roland Pennock e John W. (eds) Chapman. New York University Press, 1986.
- Ben-Ari, Mordechai, e Francesco Mondada. *Elements of Robotics*. Springer, Cham, 2018.
- Benson, Peter. «Philosophy of property law.» In *The Oxford Handbook of Jurisprudence & Philosophy of Law*, di Jules Coleman, Kenneth Einar Himma e Scott J. (eds.) Shapiro, 752—757. Oxford University Press, 2002.
- Bishop, John. *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge: Cambridge University Press , 1989.
- Blankfein-Tabachnick, David H. «Intellectual Property Doctrine and Midlevel Principles.» *California Law Review* 5, n. 101 (2013): 1315-1360.

- Boella, Guido, e Leon Van Der Torre. «A game-theoretic approach to normative multi-agent systems.» *IEEE, Transactions on Systems, Man, and Cybernetics*, 2007: 68–79.
- Böök, Lucas. «Towards a Theory of Reflexive Intentional Systems.» *Synthese* 118 (1999): 105–117.
- Braddon-Mitchell, D. «Introducing the Canberra Plan.» In *Conceptual Analysis and Philosophical Naturalism*, di D. Braddon-Mitchell e R. (eds.) Nola, 1-20. MIT Press, 2009.
- Brandom, Robert. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, 2001.
- . *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, 1994.
- Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press, 1987.
- Brown, J. P. «Toward an Economic Theory of Liability.» *Journal of Legal Studies* 2, n. 2 (1973): 323-349.
- Brożek, Bartosz, e Janika Bartosz. «Can artificial intelligences be moral agents?» *New Ideas in Psychology* 54 (2019): 101-106.
- Brunet, Pierre. «Rights of Nature and Legal Personality of Natural Entities in New Zealand: Making a Commons?» *Journal of Constitutional History* 38 (2019): 39-60.
- Bryson, J. J., M. E. Diamantis, e T. D. Grant. «Of, for, and by the people: the legal lacuna of synthetic persons.» *Artificial Intelligence and Law* 25 (2017): 273–291.
- Buckland, W. W. *The Roman Law of Slavery, The Condition of the Slave in Private Law from Augustus to Justinian*. Cambridge University Press, 2010.
- Burr, Christopher, Nello Cristianini, e James Ladyman. «An Analysis of the Interaction Between Intelligent Software Agents and Human Users.» *Minds & Machines*, 2018: 735–774.
- Calabresi, Guido. «Some Thoughts On Risk Distribution And The Law Of Torts.» *The Yale Law Journal* 70, n. 4 (1961).
- . *The Costs of Accidents; A Legal and Economic Analysis*. Yale University Press, 1997.
- Calvano, E, G Calzolari, V Denicolò, e S Pastorello. «Artificial Intelligence, Algorithmic Pricing, and Collusion.» *American Economic Review* 110, n. 10 (2020): 3267-3297.

- Canale, Damiano. «Persona.» In *Filosofia del diritto. Norme, concetti, argomenti*, di Mario Ricciardi, Andrea Rossetti e Vito (a cura di) Velluzzi. Carocci, 2015.
- Canale, Damiano, e Giovanni Tuzet. «On legal inferentialism. Toward a pragmatics of semantic content in legal interpretation?» *Ratio Juris* 20, n. 1 (2007): 32-44.
- Cane, Peter. «Mens Rea in Tort Law.» *Oxford Journal of Legal Studies* 20, n. 4 (2000): 533-556.
- Carnap, Rudolf. *Logical Foundations of Probability*. University of Chicago Press, 1950.
- Castelfranchi, C., F. Dignum, C.M. Jonker, e J. Treur. «Deliberative Normative Agents: Principles and Architecture.» In *Intelligent Agents VI. Agent Theories, Architectures, and Languages, ATAL*, *Lecture Notes in Computer Science*, di Jennings N.R. e Lespérance Y. (eds). Springer, 2000.
- Castelfranchi, Cristiano et al. *Deliberative Normative Agents: Principles and Architecture*. Berlin, Heidelberg: Springer, 2000.
- Chalmers, David J. «A Computational Foundation for the Study of Cognition.» *Journal of Cognitive Science* 12, n. 4 (2011): 323-357.
- Chalmers, David J., e Frank Jackson. «Conceptual Analysis and Reductive Explanation.» *The Philosophical Review* 110, n. 3 (2001): 315–360.
- Chen, Boyuan, Carl Vondrick, e Hod Lipson. «Visual behavior modelling for robotic theory of mind.» *Scientific Reports* 11 (2021).
- Chesterman, Simon. «Artificial Intelligence And The Limits Of Legal Personality.» *International & Comparative Law Quarterly* 69, n. 4 (2020).
- Chilovi, Samuele, e Georgios G. Pavlakos. «Law-determination as Grounding: A Common Grounding Framework for Jurisprudence.» *Legal Theory* 25 (2019): 53–76.
- Chisholm, Roderick. *Perceiving*. Ithaca: Cornell University Press, 1957.
- . *Theory of Knowledge*. Englewood Cliffs: Prentice Hall, 1977.
- Chopra, Samir, e Lawrence F. White. *A Legal Theory for Autonomous Artificial Agents*. The University of Michigan Press, 2011.
- Chopra, Samir, e Lawrence White. «Artificial Agents and the Contracting Problem: A Solution Via an Agency Analysis.» *University of Illinois Journal of Law Technology & Policy*, 2010: 363-404.
- Coase, Ronald H. «The Nature of the Firm.» *Economica* 4, n. 16 (1937): 386–405.

- Coase, Ronald H. «The Problem of Social Cost.» *The Journal of Law & Economics* 3 (1960).
- Coeckelbergh, Mark. «Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability.» *Science and Engineering Ethics* 26 (2020): 2051–2068.
- Coffee, John C. Jr. «No Soul to Damn: No Body to Kick: An Unscandalized Inquiry into the Problem of Corporate Punishment.» *Michigan Law Review* 79 (1981): 386.
- Colvin, E. «Corporate personality and criminal liability.» *Criminal Law Forum* 6 (1995): 1–44.
- Committee on Legal Affairs, Rapporteur: Stéphane Séjourné. «Intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI)).» Draft Report, 2020.
- Consiglio, Elena. «Looking for the "Vulnerable Subject": The Mencian Account of the Person.» *Rivista di filosofia del diritto, Journal of Legal Philosophy*, 2019: 143-162.
- Cooter, Robert D. «Economic Theories of Legal Liability.» *The Journal of Economic Perspectives* 5, n. 3 (1991): 11–30.
- Correia, Fabrice, e Benjamin (eds.) Schnieder. *Metaphysical Grounding: Understanding the Structure of Reality*. Cambridge University Press, 2012.
- Couch, J. A. «Woman in Early Roman Law.» *Harvard Law Review* 8 (1894): 39–50.
- Croce, Mariano. *Self-sufficiency of Law - A Critical-institutional Theory of Social Order*. Law and Philosophy Library, 2012.
- Curran, William J. «An Historical Perspective on the Law of Personality and Status with Special Regard to the Human Fetus and the Rights of Women.» *The Milbank Memorial Fund Quarterly. Health and Society* (The Milbank Memorial Fund Quarterly. Health and Society) 61, n. 1 (1983): 58-75.
- Dada, E. G., J.S., Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, e O. E. Ajibuwa. «Machine learning for email spam filtering: review, approaches and open research problems.» *Heliyon* 5, n. 6 (2019).
- Dahiyat, E.A.R. «Law and software agents: Are they “Agents” by the way?» *Artificial Intelligence and Law*, 2021: 59-86.
- Daniels, Norman. *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge University Press, 1996.

- . *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge University Press, 1996.
- Das, S, A Dey, A Pal, e N Roy. «Applications of Artificial Intelligence in Machine Learning: Review and Prospect.» *International Journal of Computer Applications*, 2015: 31- 41.
- Datta, A., S. Sen, e Y. Zick. «Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems.» *2016 IEEE symposium on security and privacy (SP)*, 2016: 598–617.
- Davidson, Donald. «Actions, Reasons, and Causes.» *The Journal of Philosophy* 60, n. 23 (1963): 685–700.
- Davies, M. «The Philosophy of Mind.» In *Philosophy 1: A Guide through the Subject*, di A. C. (ed.) Graylin, 250–335. Oxford University Press, 1998.
- Davnall, R. S. «olving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics.» *Science and Engineering Ethics* 26 (2020): 431–449.
- Deakin, Simon. «The Juridical Nature of the Firm.» In *The SAGE Handbook of Corporate Governance*, di T. Clarke and D. Branson (eds.), 113-135. Thousand Oaks, CA: Sage, 2012.
- Deakin, Simon. «Juridical Ontology: The Evolution of Legal Form.» *Historical Social Research* 40 (2015): 170-184.
- Dennett, Daniel C. «Conditions of Personhood.» In *The Identities of Persons*, di Amelie O (ed.) Rorty. University of California Press, 1976.
- Dennett, Daniel. *Kind of Minds*. Basic Books, 2008.
- . *The Intentional Stance*. Cambridge: MIT Press, 1987.
- Dewey, John. «The Historic Background of Corporate Legal Personality.» *Yale Law Journal* 35 (1926): 655-673.
- Diakopoulos, Nicholas. «Accountability, Transparency and Algorithms.» In *The Oxford Handbook of the Ethics of AI*, di Markus Dubber, Frank Pasquale e Sunit (eds) Das, 197-213. Oxford University Press, 2020.
- Diaz-Leon, E. «Reductive explanation, concepts, and a priori entailment.» *Philosophical Studies* 155 (2011): 99–116.
- Dignum, Frank. «Autonomous agents with norms.» *Artificial Intelligence and Law* 7 , n. 1 (1999): 69–79.
- Dombrowski, Daniel. *Babies and Beasts: The Argument from Marginal Cases*. The University of Illinois Press, 1997.

- Dretske Fred, in. «The intentionality of cognitive states.» *Midwest Studies in Philosophy* (Oxford: Oxford University Press) 5, n. 1 (1980): 281-294.
- Dreyfus, Hubert L. *What Computers Still Can't Do. A Critique of Artificial Reason*. MIT Press, 1992.
- Duff, Anthony R. *Answering for crime: Responsibility and liability in the criminal law*. Bloomsbury Publishing, 2007.
- Duff, Antony R. *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law*. Blackwell, 1990.
- Dworkin, Ronald. *Law's Empire*. Harvard University Press, 1986.
- Edwards, J.L.J. «The Criminal Degrees Of Knowledge.» *The Modern Law Review* 17 (1954): 293-314.
- Endsley, M. R., e D. M. Kraber. «Level of automation effects on performance, situation awareness and workload in a dynamic control task.» *Ergonomics*, 1999: 462–492.
- Epstein, Brian. *Social Ontology*, E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* . 2018 .
<https://plato.stanford.edu/archives/sum2018/entries/social-ontology/>.
- . *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford University Press, 2015.
- Epstein, Brian. «A Framework for Social Ontology.» *Philosophy of the Social Sciences* 46, n. 2 (2016): 147–167.
- Epstein, Brian. «Anchoring versus Grounding: Reply to Schaffer.» *Philosophy and Phenomenological Research* 99, n. 3 (2019).
- Epstein, Brian. «How Many Kinds of Glue Hold the Social World Together?» In *Perspectives on Social Ontology and Social Cognition: Studies in the Philosophy of Sociality*, di Mattia Gallotti e John Michael. Springer, 2014.
- Epstein, Brian. «Replies to Hawley, Mikkola, and Hindriks.» *Inquiry: An Interdisciplinary Journal of Philosophy* 62, n. 2 (2019): 230-246 .
- Evans, Edward Payson. *The Criminal Prosecution and Capital Punishment of Animals*. London William Heinemann, 1906.
- Ferraris, Maurizio. *Documentality. Why It Is Necessary to Leave Traces*. Fordham University Press, 2012.
- Fine, Kit. «Guide to Ground.» In *Metaphysical Grounding: Understanding the Structure of Reality*, di F. Correia e B. Schneider, 37–80. Cambridge University Press, 2012.

- Fine, Kit. «Ontological Dependence.» *Proceedings of the Aristotelian Society* 95 (1995): 269–290.
- Fine, Kit. «Vagueness, Truth and Logic.» *Synthese* 30 (1975): 265–300.
- Fineman, Martha A. «The Vulnerable Subject: Anchoring Equality in the Human Condition.» *Yale Journal of Law & Feminism* 20, n. 1 (2008): 1-23.
- Fischer, J. «Computers as agents: a proposed approach to revised .» *Indiana Law Journal* 72, n. 2 (1997): 545–570.
- Fischer, John Martin,, e Mark Ravizza. *Responsibility and control*. Cambridge University Press, 2000.
- Fodor, Jerry. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press, 1998.
- . *The Language of Thought*. New York: Thomas Y. Crowell, 1975.
- Fodor, Jerry, e Ernest Lepore. *The Compositionality Papers*. Oxford University Press, 2002.
- Fodor, Jerry, e Ernest Lepore. «The red herring and the pet fish: why concepts still can't be prototypes.» *Cognition* 58, n. 2 (1996): 253-270.
- Foot, Philippa. «The Problem of Abortion and the Doctrine of the Double Effect.» *The Oxford Review*, 1967.
- Foucault, Michel. *Surveiller et Punir*. 1st edition. Gallimard, 1975.
- Frase, R. S. «Punishment purposes.» *Stanford Law Review* 58, n. 1 (2005): 67-84.
- Freund, Ernst. *The Legal Nature of Corporations*. Batoche Books , 2000.
- Frohm, Jörgen, Veronica Lindström, Mats Winroth, e Johan Stahre. «Levels of Automation in Manufacturing.» *Ergonomia. International Journal of Ergonomics and Human Factors*, 2008: 181-207.
- Galgano, Francesco. *Le insidie del linguaggio giuridico. Saggio sulle metafore nel diritto*. Il Mulino, 2010.
- . *I Fatti Illeciti*. CEDAM, 2008.
- Gardner, J. F. *Women in Roman Law and Society*. Routledge, 1987.
- Gardner, John. «The Mark of Responsibility.» *Oxford Journal of Legal Studies* 23, n. 2 (2003): 157-171.
- Gellers, Joshua C. *Rights for Robots Artificial Intelligence, Animal and Environmental Law*. Routledge, 2021.
- Gerber, David. «Strict Criminal Liability and Justice.» *Archives for Philosophy of Law and Social Philosophy* 60, n. 4 (1974): 513- 545.
- Gettier, Edmund L. «Is Justified True Belief Knowledge?» *Analysis* 23, n. 6 (1963): 121–123.

- Gibert, M., e D. Martin. «In search of the moral status of AI: why sentience is a strong argument.» *AI & Society*, 2021.
- Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, e L. Kagal. «Explaining Explanations: An Overview of Interpretability of Machine Learning.» *The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018)*. 2018. 80-89.
- Gindis, D. «Legal personhood and the firm: Avoiding anthropomorphism and equivocation.» *Journal of Institutional Economics* 12, n. 3 (2016).
- Gindis, David. «From Fictions and Aggregates to Real Entities in the Theory of the Firm.» *Journal of Institutional Economics* 5, n. 1 (2009): 25–46.
- Goodman, Nelson. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.
- Gordon, John Stewart. «Artificial moral and legal personhood.» *AI & Society* 36, n. 2 (2021): 1-15.
- Gordon, John Stewart. «Artificial moral and legal personhood.» *AI & Society*, 2020.
- Gray, John Chipman. *The nature and sources of the law*. Adamant Media Corporation, 2006.
- . *The Nature and Sources of the Law*. Macmillan, 1921.
- Greif, A., e C. Kingston. «Institutions: Rules or Equilibria?» In *Political Economy of Institutions, Democracy and Voting*, di N. Schofield e G. (eds.) Caballero. Springer, 2011.
- Greif, A., e C. Kingston. «Institutions: Rules or Equilibria?» In *olitical Economy of Institutions, Democracy and Voting*, di N. Schofield and G. Caballero (eds.). Springer,, 2011.
- Guala, Francesco. «Epstein on Anchors and Grounds.» *Journal of Social Ontology*, 2 (2016).
- Guala, Francesco. «The Philosophy of Social Science: Metaphysical and Empirical.» *Philosophy Compass* 2 (2007): 954-980.
- . *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press, 2016.
- Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, e Dino Pedreschi. «A survey of methods for explaining black box models.» *ACM Computing Surveys* 51, n. 5 (2018): 1-42.

- Guoqiang, H., T. Wee Peng, e W. Yonggang. «Cloud Robotics: Architecture, Challenges and Applications.» *NETWORK*, 2012: 21-28.
- Hacker, Philipp, Ralf Krestel, Stefan Grundmann, e Felix Naumann. «Explainable AI under contract and tort law: legal incentives and technical challenges.» *Artificial Intelligence and Law* 28 (2020): 415–439.
- Hage, Jaap. «The Meaning of Legal Status Words.» In *Concepts in Law*, di Jaap Hage e Dietmar (eds) von der Pfordten, 55-66. Springer, 2009.
- Hallevy, Gabriel. *Liability for Crimes Involving Artificial Intelligence Systems*. Springer, 2014.
- Hallevy, Gabriel. «The criminal liability of artificial intelligence entities. From science fiction to legal social control.» *Akron Intellectual Property Journal*, 2010.
- Hansmann, Henry, Reinier Kraakman, e Richard [] 1335 Squire. «Law and the Rise of the Firm.» *Harvard Law Review* 119, n. 5 (2005): 1333-1403.
- Hart, Herbert L. A. «Definition and Theory in Jurisprudence.» In *Essays in Jurisprudence and Philosophy*, di Herbert L. A. Hart. Oxford University Press, 1984.
- Hayek, Friederich Von. *Law, legislation and liberty, vol 1: Rules and order*. University of Chicago Press, 1973.
- He, X., K. Zhao, e X. Chu. «AutoML: A survey of the state-of-the-art.» *Knowledge-Based Systems* 212 (2021).
- Henley, Kenneth. «Abstract Principles, Mid-Level Principles, and the Rule of Law.» *Law and Philosophy* 12, n. 1 (1993): 121-132.
- Hildebrandt, Mirelle. «From Galatea 2.2 to Watson – and back?» In *Human Law and Computer Law: Comparative Perspectives*, di Mireille Hildebrandt e Jeanne (eds) Gaakeer, 23-45. Springer Nature, 2013.
- Himma, Kenneth Einar. «Conceptual Jurisprudence. An Introduction to Conceptual Analysis and Methodology in Legal Theory.» *Revus* 26 (2015).
- Hindriks, Frank. *Rules & Institutions: Essays in Meaning, Speech Acts and Social Ontology*. Haveka B. V, 2005.
- Hindriks, Frank, e Francesco Guala. «Institutions, rules, and equilibria: a unified theory.» *Journal of Institutional Economics* 11, n. 3 (2015): 459 – 480.

- Hohfeld, Wesley N. «Some Fundamental Legal Conceptions as Applied in Judicial Reasoning,» *Yale Law Journal* 23 (1913): 16–59.
- Huettinger, Maik, e Jonathan Andrew Boyd. «Taxation of robots – what would have been the view of Smith and Marx on it?» *International Journal of Social Economics* 47, n. 1 (2019): 41–53.
- Huff, T. *The Rise of Early Modern Science. Islam, China and the West*. Cambridge University Press, 2003.
- Hume, David. *A Treatise of Human Nature*. 1778. A cura di L. A. Selby-Bigge e P. H. Nidditch. Oxford University Press, 1740.
- Jackson, Frank. *From Metaphysics to Ethics - A defence of Conceptual Analysis*. Oxford University Press, 2000.
- Jackson, Frank, Robert Pargetter, e Elizabeth Prior. «Functionalism and Type-Type Identity Theories.» *Philosophical Studies* 42 (1982): 209–225.
- Jhering, Rudolf. *Der Geist des römischen Rechts auf den verschiedenen Stufen seiner Entwicklung*. Leipzig: Breitkopf und Härtel, 1852.
- Keeton, G W. *The Elementary Principles of Jurisprudence*. Sir Isaac Pitman and Sons, 1930.
- Kelsen, Hans. *General Theory Of Law And The State*. Traduzione di Anders Wedberg. Harvard University Press, 1945.
- Kerr, I. R. «Spirits in the Material World: Intelligent Agents as Intermediaries in Electronic Commerce.» *Dalhousie Law Journal* 22 (1999): 189-249.
- Khalidi, Muhammad Ali. «Kinds: Natural vs. Human Kinds.» In *Encyclopedia of Philosophy and the Social Sciences*, di B. (ed) Kaldis. Thousand Oaks, CA: Sage, 2013.
- Khalidi, Muhammad Ali. «Law as a social kind.» *Proceedings of the Joint Ontology Workshops*, 2019.
- Khalidi, Muhammad Ali. «Mind-Dependent Kinds.» *Journal of Social Ontology* 2 (2016): 223–246.
- Khalidi, Muhammad Ali. «Natural kinds as nodes in causal networks.» *Synthese* 195 (2018): 1379-1396.
- Khalidi, Muhammad Ali. «Three Kinds of Social Kinds.» *Philosophy and Phenomenological Research* 90 (2015): 96-112.
- Kingsbury, J., e J. McKeown-Green. «Jackson’s armchair: The only chair in town?» In *Conceptual Analysis and Philosophical Naturalism*, di D. Braddon-Mitchell e R. (eds) Nola, 159-182. MIT Press, 2009.
- Kramer, Matthew. «Some Doubts about Alternatives to the Interest Theory.» *Ethics*, 2013: 245–263.

- Krasheninnikova, E, J García, R Maestre, e F Fernández. «Reinforcement learning for pricing strategy optimization in the insurance industry.» *Engineering Applications of Artificial Intelligence* 80 (2019): 8-19.
- Kriegel, Uriah. «Interpretation: Its Scope and Limits.» In *New Waves in Metaphysics. New Waves in Philosophy*, di Allan (ed) Hazlett. Palgrave Macmillan,, 2010.
- Kriegel, Uriah. «Is Intentionality Dependent upon Consciousness?» *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 116, n. 3 (2003): 271-307.
- Kroll, A. Joshua. «Accountability in Computer Systems.» In *The Oxford Handbook of the Ethics of AI*, di Markus Dubber, Frank Pasquale e Sunit (eds) Das, 180-196. Oxford University Press, 2020.
- Kurki, Visa. *A Theory of Legal Personhood*. Oxford Legal Philosophy, 2019.
- Kurki, Visa, e Tomasz Pietrzykowski. *Legal Personhood: Animals, Artificial Intelligence and the Unborn*. A cura di Visa Kurki e Tomasz Pietrzykowski. Springer International Publishing, 2017.
- Kurzweil, Raymond. *The singularity is near*. Viking press, 2005.
- La Torre, Massimo. «Esistenzialismo e Istituzionalismo.» In *Che cosa è il diritto, Ontologie e concezioni del giuridico*, di Giorgio Bongiovanni, Giorgio Pino e Corrado Roversi, 115-147. Giappichelli Editore, 2016.
- . *Law as Institution*. Springer, Law and Philosophy Library, 2010.
- Lackoff, George, e Mark Johnson. *Metaphors we live by*. The University of Chicago Press, 2003.
- Lagioia, Francesca, e Giovanni Sartor. «AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective.» *Philosophy & Technology* 33 (2020): 433–465.
- Landes, William M, e Richard Posner. *The Economic Structure of Tort Law*. Harvard University Press, 1987.
- Larson, J, S Mattu, L Lauren Kirchner, e J Angwin. «How We Analyzed the COMPAS Recidivism Algorithm.» *ProPublica*. 23 May 2016. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- Laskowski, N. G., e Stephen Finlay. «Conceptual Analysis in Metaethics.» In *The Routledge Handbook of Metaethics*, di T. McPherson e Plunkett David (eds.), 536-551. Routledge, 2017.

- Latorre, Massimo. «Ota Weinberger, Neil MacCormick e il neoistituzionalismo giuridico.» In *Filosofi del diritto contemporanei*, di Gianfrancesco Zanetti, 1-30. Raffele Cortina Editore, 1999.
- Laukyte, M. «Artificial agents among us: Should we recognize them as agents proper?» *Ethics and Information Technology*, 2017: 1–17.
- Lawson, Frederik H. «The Creative Use of Legal Concepts.» *New York University Law Review* 32 (1957).
- Leary, Stephanie. «Normativity.» In *The Routledge Handbook of Metaphysical Grounding*, di M. Raven. Routledge Handbooks in Philosophy, 2020.
- Leigh, L. H. «The Criminal Liability of Corporations and Other Groups: A Comparative View.» *Michigan Law Review* 80, n. 7 (1982): 1508–1528.
- List, Christian. «Group Agency and Artificial Intelligence.» *Philosophy and Technology* 4 (2021): 1-30.
- List, Christian, e Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011.
- Lorentzen, K.F. «Luhmann Goes Latour: Zur Soziologie hybrider Beziehungen.» In *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*, di W. Rammert e I. (eds.) Schulz-Schaeffer. 2002.
- Lorini, Giuseppe. «Meta-Institutional Concepts: A New Category for Social Ontology.» *Rivista di estetica*, 2014: 127-139.
- Lorini, Giuseppe, e Wojciech Żelaniec. «The Background of Constitutive Rules: Introduction.» *Argumenta (Special Issue)* 13, n. 4 (2018): 9-19.
- Luhmann, Nicholas. *Social Systems*. Stanford University Press, 1996.
- M. La Torre, Esistenzialismo e Istituzionalismo, in G. Bongiovanni, G. Pino, C. Roversi, Che cosa è il diritto, Ontologie e concezioni del giuridico, Giappichelli Editore, 2016, pp. 115-147. s.d.
- MacCarthy, John. *Professor John McCarthy, What is Artificial Intelligence?.* 2007. <http://jmc.stanford.edu/articles/whatisai.html>.
- MacCormick, Neil. «Further Thoughts on Institutional Facts.» *Revue internationale de semiotique juridique* 5 (1992): 3–15.
- MacCormick, Neil. «Law as an institutional fact.» In *An Institutional Theory of Law, New Approaches to Legal Positivism*, di Neil MacCormick e Ota Weinberger. Springer, 1986.
- . *Institutions of Law, An Essay in Legal Theory*. Oxford University Press, 2007.

- MacCormick, Neil. «Institutions, Arrangements and Practical Information.» *Ratio Juris* 1, n. 1 (1988): 73-82.
- MacCormick, Neil. «Norms, Institutions, and Institutional Facts.» *Law and Philosophy* 17 (1998).
- MacCormick, Neil. «Persons as Institutional Facts.» In *Reine Rechtslehre im Spiegel ihrer Fortsetzer und Kritiker*, di Ota Weinberger e Werner (eds) Krawietz. Springer, 1988.
- MacCormick, Neil, e Ota Weinberger. *An Institutional Theory of Law, New Approaches to Legal Positivism*. Springer, 1986.
- Macey, Jonathan, e Joshua Mitts. «Finding Order in the Morass: The Three Real Justifications for Piercing the Corporate Veil.» *Cornell Law Review* 100, n. 1 (2014): 99-155.
- Marchesi, S., D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, e A. Wykowska. «Do We Adopt the Intentional Stance Toward Humanoid Robots?» *Frontiers in Psychology*, 2019.
- Margolis, E., e S. Laurence. "Concepts", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). 2021. <https://plato.stanford.edu/archives/spr2021/entries/concepts/>.
- Margolis, Eric, e Stephen Laurence. «Concepts.» *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). spring edition 2021 . <<https://plato.stanford.edu/archives/spr2021/entries/concepts/>>.
- Marmor, Andrei. «Deep Conventions.» *Philosophy and Phenomenological Research* 74, n. 3 (2007): 586-610.
- Matthias, Andreas. *Automaten als Träger von Rechten. Plädoyer für eine Gesetzänderung*. Logos Verlag, 2008.
- Meckling, William, e Michael Jensen. «Reflections on the Corporation as a Social Invention.» *Midland Corporate Finance Journal* 1, n. 3 (1983): 6–15.
- Mikkola, Mari. «Grounding and Anchoring: On the Structure of Epstein's Social Ontology.» *Inquiry* 198 (2017).
- Miller, Dolores. «Constitutive Rules and Essential Rules.» *Philosophical Studies* 39 (1981): 183-197.
- Minsky, Marvin. *The Society of Mind*. Cambridge: MIT Press, 1986.
- Mitchell, Tom M. *Machine Learning*. McGraw Hill Education, 1997.
- Moore, Michael S. *Law and Psychiatry: Rethinking the Relationship*. Cambridge University Press, 1984.

- Moreso, Jose Juan, e Chiara Valentini. «In the Region of Middle Axioms: Judicial Dialogue as Wide Reflective Equilibrium and Mid-level Principles.» *Law and Philosophy* 40, n. 5 (2021): 1-39.
- Morocho-Cayamcela, M. E., L. Haeyoung, e L. Wansu. «Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions.» *IEEE Access*, 2019.
- Naffine, Ngaire. *The Law's Meaning of Life, Philosophy, Religion, Darwin and the Legal Person*. Hart Publishing, 2009.
- Naffine, Ngaire. «Who are Law's Persons? From Cheshire Cats to Responsible Subjects.» *Modern Law Review* 66, n. 3 (2003): 346–367.
- Nagel, Jennifer. *Knowledge: A Very Short Introduction*. Oxford University Press, 2014.
- Nagel, Thomas. «What is it like to be a bat?» *The Philosophical Review* 83, n. 4 (1974): 435-450.
- Nékam, Alexander. *The Personality Conception of the Legal Entity*. Cambridge: Harvard University Press, 1938.
- Nolan, Daniel. «Platitudes and metaphysics.» In *Conceptual Analysis and Philosophical Naturalism*, di D. Braddon-Mitchell e R. (eds.) Nola. MIT Press, 2009.
- North, Douglas. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1990.
- Nozick, Robert. *The Nature of Rationality*. Princeton University Press, 1993.
- Nussbaum, Martha. *Women and Human Development: The Capabilities Approach*. Cambridge University Press, 2000.
- Nuti, G., M. Mirghaemi, P. Treleven, e C. Yingsaeree. «Algorithmic Trading.» *Computer* 44, n. 11 (2011): 61–69.
- Oberson, Xavier. *Taxing Robots? Helping the Economy to Adapt to the Use of Artificial Intelligence*. Edward Elgar Pub, 2019.
- Oh, S.J., B. Schiele, e M. Fritz. «Towards reverse-engineering black-box neural networks.» In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, di W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen e K. R. (eds) Müller. Springer, 2019.
- Omohundro, Stephen M. «The basic AI drives.» *Artificial General Intelligence, Frontiers in Artificial Intelligence and Applications* 171 (2008): 483-492.

- Ostrom, E. *Understanding Institutional Diversity*. Princeton University Press, 2005.
- Parasuraman, R., T. B. Sheridan, e C. D. Wickens. «A model for types and levels of human interaction with automation.» *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 30.3, 2000: 286–297.
- Parikh, Rahul. «Social Software.» *Synthese* 132 (2002): 187–211.
- Parsons, Kathryn Pyne. «Three Concepts of Clusters.» *Philosophy and Phenomenological Research*, 33, n. 4 (1973): 514-523.
- Patel, T. A., et al. «Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods.» *Cancer* 123, n. 1 (2016): 114-121.
- Peacocke, Christopher. *A Study of Concepts*. MIT Press, 1992.
- Pelletier, Francis Jeffrey. «The Principle of Semantic Compositionality.» *Topoi* 13 (1994): 11–24.
- Pietrzykowski, Tomasz. « The Idea of Non-personal Subjects of Law.» In *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, di Visa A. J. Kurki e Tomasz (eds) Pietrzykowski. Springer International Publishing, 2017.
- . *Personhood Beyond Humanism Animals, Chimeras, Autonomous Agents and the Law*. Springer, 2018.
- Pinker, Steven. *How the Mind Works*. Princeton University Press, 1997.
- Plunkett, David. «A Positivist Route for Explaining How Facts Make Law.» *Legal Theory* 139 (2012).
- Plunkett, David. «Expressivism, representation, and the nature of conceptual analysis.» *Philosophical Studies* 156, n. 1 (2011): 15-31.
- Poste, Edward. *Gains Institutiones or Institutes of Roman Law*. Clarendon Press, 1904.
- Psillos, Stathis. « Rudolf Carnap’s ‘theoretical concepts in science.» *Studies in History and Philosophy of Science*, 2000: 151–172.
- Pütz, F., F. Murphy, M. Mullins, K. Maier, R. Friel, e T. Rohlf. «Reasonable, Adequate and Efficient Allocation of Liability Costs for Automated Vehicles: A Case Study of the German Liability and Insurance Framework.» *European Journal of Risk Regulation*, 2018: 548 - 563.
- Rahwan, Iyad. «Society-in-the-loop: programming the algorithmic social contract.» *Ethics and Information Technology* 20 (2018): 5–14.
- Ramsey, Frank. *The foundations of mathematics and other essays*. London: Routledge, 1931.

- Rasmusen, E. «Agency Law and Contract Formation.» *American Law and Economics Review* 6, n. 2 (2004): 369–409.
- Rawls, John. *A Theory of Justice*. Harvard University Press, 1971.
- Rawls, John. «The Independence of Moral Theory.» *Proceedings and Addresses of the American Philosophical Association* 47 (1974): 5–22.
- Raz, Joseph. *Ethics in the Public Domain*. Oxford University Press, 1995.
- Renner, K. *The Institutions of Private Law and Their Social Functions*. Routledge & Kegan Paul Limited, 1949.
- Ribeiro, M. T., S. Singh, e C. Guestrin. «Why should i trust you? Explaining the predictions of any classifier.» *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,. 2016. 1135–1144.
- Ribeiro, M.T., S. Singh, e C. Guestrin. «Anchors: High-precision model-agnostic explanations.» *AAAI Conference on Artificial Intelligence*. 2018. 1527–1535.
- Rosen, Gideon. «Ground by Law.» *Philosophical Issues* 27 (2017): 279-301.
- Rosen, Gideon. «Metaphysical Dependence: Grounding and Reduction.» In *Modality: Metaphysics, Logic, and Epistemology*, di Bob Hale e Aviv (eds) Hoffmann, 109-135. 2010.
- Rosen, Gideon. «Real Definition.» *Analytic Philosophy* 56, n. 3 (2015).
- Rosen, Gideon. «What is a Moral Law?» *Oxford Studies in Metaethics* 12 (2017).
- Ross, Alf. «Definition in Legal Language.» *The Journal of Symbolic Logic* 25 (1960).
- Ross, Alf. «Tu[^]-Tu[^].» *Scandinavian Studies in Law* 1 (1957): 139–153.
- Roversi, Corrado. «A three-dimensional ontology of customs.» In *Spaces of Law and Custom*, di Edoardo Frezet, Marc Goetzmann e Luke Mason. Routledge, 2021.
- Roversi, Corrado. «Conceptualizing Institutions.» *Phenomenology and the Cognitive Sciences* 13 (2014): 201–215.
- . *Costituire: Uno studio di ontologia giuridica*. Giappichelli Editore, 2012.
- Roversi, Corrado. «In defence of constitutive rules.» *Synthese*, 2021.
- Rudin, Cynthia. «Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.» *Nature Machine Intelligence*, 2019.
- Ruiter, Dick. *Legal Institutions*. Kluwer Academic Publishers, 2001.
- Ruiter, Dick W. P. *Institutional Legal Facts: Legal Powers and their Effects*. Kluwer Academic Publishers , 1993.

- Russell, Stuart J., e Peter Norvig. *Artificial Intelligence, A Modern Approach*. Global Edition. Pearson Education, 2016.
- Sartor, Giovanni. «Cognitive automata and the law: Electronic contracting and the intentionality of software agents.» *Artificial Intelligence and Law*, 2009 : 253 - 290.
- Sartor, Giovanni. «Gli agenti software: nuovi soggetti di cyberdiritto.» *Contratto e Impresa* 2 (2002): 57-91.
- Sartor, Giovanni. «Legal concepts as inferential nodes and ontological categories.» *Artificial Intelligence and Law* 17 (2009): 217–251.
- Schaffer, Jonathan. «Anchoring as Grounding: On Epstein’s The Ant Trap.» *Philosophy and Phenomenological Research* 99 (2019): 749–767.
- Schaffer, Jonathan. «On what grounds what.» In *Metametaphysics: New Essays on the Foundations of Ontology*, di David Manley, David J. Chalmers e Ryan (eds.) Wasserman, 347-383. Oxford University Press, 2009.
- Schwarz, Gustav. «Rechtssubjekt u. Rechtszweck.» *Archiv für bürgerliches Recht* 32 (1908): 1-24.
- Schwyzler, Hubert. «Rules and Practices.» *The Philosophical Review* 78, n. 4 (1969): 451-467.
- Searle, John. *Making the Social World, The Structure of Human Civilization*. Oxford University Press, 2010.
- . *Mind: A Brief Introduction*. Oxford University Press, 2004.
- Searle, John R. «Constitutive Rules.» *Argumenta* 4, n. 1 (2018).
- . *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.
- Searle, John. *The Construction of Social Reality*. The Free Press, 1996.
- . *The Rediscovery of the Mind*. MIT Press, 1992.
- Sellars, Wilfrid. «Inference and Meaning.» *Mind* 62 (1953): 313-338.
- Sharkey, Amanda. «Autonomous weapons systems, killer robots and human dignity.» *Ethics and Information Technology* 21 (2019): 75–87.
- Silva, Denis Franco. «From Human to Person: Detaching Personhood from Human Nature.» In *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, di Visa A J Kurki e Tomasz Pietrzykowski, 113-125. Springer International Publishing, 2017.
- Silver, D., A. Huang, C. Maddison, e et al. «Mastering the Game of Go with Deep Neural Networks and Tree Search.» *Nature* 529 (2016): 484–489.
- Simon, H. A. «A Behavioral Model of Rational Choice.» *Quarterly Journal of Economics* 69, n. 1 (1955): 99-118.

- Simons, Kenneth W. «When Is Strict Criminal Liability Just?» *The Journal of Criminal Law and Criminology* 87, n. 4 (1997): 1075 - 1137.
- Singer, Peter. *Animal Liberation*. Harper Collins, 1975.
- Smed, S. «Intelligent software agents and agency law.» *Santa Clara Computer High Technology Law Journal* 14 (1998): 503–507.
- Solaiman, Sheikh Mohammad. « Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy.» *Artificial Intelligence and Law* 25, n. 2 (2017): 155 –179.
- Solum, Lawrence B. «Legal Personhood for Artificial Intelligences.» *North Carolina Law Review* 70, n. 4 (1992).
- Stein, P. *Roman Law In European History*. Cambridge University Press, 1999.
- Stone, Christopher D. «Should trees have standing? Toward legal rights for natural objects.» *Southern California Law Review* 45 (1972): 450-501.
- Street, T. A. *The Theory and Principles of Tort Law (Foundations of Legal Liability)*. Beard Books, 1999.
- Sunstein, Cass. *Legal Reasoning and Political Conflict*. Oxford University Press, 1996.
- Tanner, Julia K. “ ¶ 18 51. «The Argument from Marginal Cases and the Slippery Slope Objection.» *Environmental Values* 18, n. 1 (2009): 51-66.
- Teubner, Gunther. « Rights of non-humans? Electronic agents and animals as new actors in politics and law.» *Journal of Law and Society*, 2006: 497-521.
- Teubner, Gunther. «Digitale Rechtssubjekte? Zum privatrechtlichen Status autonomer Softwareagenten.» *Archiv fuer die civilistische*, 2018: 155-205.
- Thomasson, Amie. «Realism and Human Kinds.» *Philosophy and Phenomenological Research* 67 , n. 3 (2003).
- Trahan, J.R. «The Distinction Between Persons & Things: An Historical Perspective.» *Journal of Civil Law Studies* 1 (2008).
- Trist, E, e H. (ed) Murray. *The Social Engagement of Social Science*. Vol. Volume II: The Socio-Technical Perspective. Philadelphia: University of Pennsylvania, 1993.
- Tucnik, Petr. «Automatic Trading System Design.» In *Pervasive Computing for Business: Trends and Applications*, di Varuna (ed.) Godara. Sydney: IGI Global, 2010.
- Tur, Richard. *The 'Person' in Law*. Basil Blackwell, 1987.

- Umbrello, Steven, Angelo Frank De Bellis, e Phil Torres. «The future of war: could lethal autonomous weapons make conflict more ethical?» *Artificial Intelligence & Society* 35 (2020): 273–282.
- Van Duffel, Siegfried. «In Defense of the Will Theory of Rights.» *Res Publica* 18 (2012): 321–331.
- Véliz, Carissa. «Moral zombies: why algorithms are not moral agents.» *AI & Society*, 2021.
- Vernon, V. «The Coming Technological Singularity: How to Survive in the Post-Human Era.» *Vision-21, Interdisciplinary Science and Engineering in the Era of Cyberspace, Proceedings of a symposium cosponsored by the NASA Lewis Research Center and the Ohio Aerospace Institute*, 1993: 11-22.
- von der Pfordten, Dietmar. «About Concepts in Law.» In *Concepts in Law.*, di Hage Jaap C. e von der Pfordten Dietmar (eds), 17-33. Springer, 2009.
- Wachter, S., B. Mittelstadt, e L. Floridi. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.» *International Data Privacy Law* 7, n. 2 (2017): 76–99.
- Wagner, Gerhard. «Comparative Tort Law.» In *The Oxford Handbook of Comparative Law*, di Mathias Reimann e Reinhard (eds) Zimmermann, 1004 – 1040. 2012.
- Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Harvard University Press, 1994.
- Wang, Pei. *Rigid Flexibility The Logic of Intelligence*. Springer Netherlands, 2006.
- Wang, Pei. «What Do You Mean by “AI”?» *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 2008: 362-373.
- Wang, Pei, Kai Liu, e Quinn Dougherty. «Conceptions of Artificial Intelligence and Singularity.» *Information* 79, n. 9 (2018).
- Watson, A. *The Law of Persons in the Later Roman Republic*. Oxford University Press, 1967.
- Weinberger, Ota. *Ontologie, Hermeneutik und der Begriff des geltenden Rechts*. Rechtsgeltung, 1986.
- Wendehorst, Christiane. «Strict Liability for AI and other Emerging Technologies.» *Journal of European Tort Law* 11, n. 2 (2020): 150-180.

- Wheeler, Michael. «Autonomy.» In *The Oxford Handbook of the Ethics of AI*, di Markus D. Dubber, Frank Pasquale e Sanit and Das, 343-357. Oxford University Press, 2020.
- Wise, Steven. *Rattling the Cage: Toward Legal Rights For Animals*. Da Capo Press, 2001.
- . *Rattling the Cage: Towards Legal Rights for Animals*. Perseus Publishing, 2000.
- Wittgenstein, Ludwig. *Philosophical Investigation*. MacMillan, 1953.
- Wood, Luke J., Abolfazl Zaraki, Ben Robins, e Kerstin Dautenhahn. «Developing Kaspar: A Humanoid Robot for Children with Autism.» *International Journal of Social Robotics* 13 (2021): 491–508.
- Wooldridge, Michael, e Nick Jennings. «Intelligent agents: Theory and practice.» *The Knowledge Engineering Review*, 1995: 115 - 152.
- Wu, Stephen. «Autonomous vehicles, trolley problems, and the law.» *Ethics and Information Technology* 22 (2020): 1–13.
- Zadeh, Lotfi A. «Fuzzy sets.» *Information and Control* 8 (1965): 338-353.
- Zatti, Paolo. *Persona giuridica e soggettività: per una definizione del concetto di persona nel rapporto con la titolarità delle situazioni soggettive*. Padova: CEDAM, 1975.
- Zech, Herbert. «Liability for AI: public policy considerations.» *ERA Forum* 22 (2021): 147–158.
- Zhang, Q. «Fuzziness-vagueness-generality-ambiguity.» *Journal of Pragmatics* (Elsevier Science) 29 (1998): 13-31.
- Zhihui, D., H. Ligang, C. Yinong, Xiao Y., G. Peng, e W. Tongzhou. «Robot Cloud: Bridging the power of robotics and cloud computing.» *Future Generation Computer Systems*, 2017: 337-348.
- Ziemianin, Karolina. «Civil legal personality of artificial intelligence. Future or utopia?» *Internet Policy Review* 10, n. 2 (2021): 1-22.
- Zwetsloot, Remco, e Allan Dafoe. «Thinking About Risks From AI: Accidents, Misuse and Structure.» *Lawfare*. 11 February 2019.