Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

PHILOSOPHY, SCIENCE, COGNITION, AND SEMIOTICS (PSCS)

Ciclo 33

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

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

INTEGRATED INFORMATION THEORY: AN EMPIRICALLY TESTABLESOLUTION TO THE MIND-BODY PROBLEM

Presentata da: Francesco Ellia

Coordinatore Dottorato

Claudio Paolucci

Supervisore

Francesco Bianchini

Co-supervisore

Giulio Tononi

Esame finale anno 2021

Abstract of the Thesis

The purpose of this dissertation is to show that consciousness can be studied scientifically and to illustrate a possible way to do it. In particular, I aim to show how Integrated Information Theory (IIT) can provide an empirically testable solution to the mind-body problem. In the first chapter, I introduce the mind-body problem, as the general problem of how mind and matter are related, and the hard problem, as the more specific problem of explaining how consciousness arises from matter. I then proceed to illustrate the terminology and work definitions that I will employ throughout the entire work. In the second chapter, I look at the mind-body problem from a historical perspective. In particular, I go through Galileo's influential distinction between primary properties and secondary qualities that can be found in his *The Assayer*. In particular, I argue that what alienates consciousness from objective science is the ill conceptualization of the primary-properties and secondary-qualities distinction. In the next chapter, I illustrate IIT and its axiomatic approach. After an overview of IIT, I discuss what can be called the 'core' of the theory: its axioms, its postulates, and its fundamental identity, and I discuss its explanatory structure. Finally, in the last chapter, I engage with the unfolding argument, which has been presented recently as a confutation of IIT. In my analysis, I show how the unfolding argument presents several criticalities. In particular, I argue that the unfolding argument commits to a radical version of functionalism that is unfit to deal with the challenges posed by assessing consciousness in a clinical context. Finally, I show how the notion of science implied by the proponents of the unfolding argument is too strict to be useful to any analysis and it does not consider the debate on the demarcation problem in philosophy of science.

Table of Contents

Acknowledgments

I desire to express my sincere gratitude to all the people who helped me during my Ph.D. In particular I am grateful for the opportunity that Francesco Bianchini gave me to become his student and work under his supervision. His precious insights and generous encouragements were instrumental for my growth as a researcher. I am also grateful to Francesco for giving me the opportunity to teach in his classes, which was one of the most fulfilling experiences of the last three years. I want to thank Raffaella Campaner for the support she gave me, from helping me to delineate important aspects of my research project to encouraging me to pursue it abroad. I also wish to thank Sebastiano Moruzzi for his help and support in general and in particular during the seminars of the European Ph.D. Network in Philosophy. My gratitude also goes to Marco Viola for his advice, especially in the early stages of my program.

My deepest gratitude goes also to both Giulio Tononi and Chiara Cirelli for welcoming me at their lab at the University of Wisconsin-Madison. Studying at the Center for Sleep and Consciousness changed radically my perspective on science and philosophy, and in general was an incredibly formative experience. I am particularly grateful to Professor Tononi for all the time he dedicated to me and to my questions. His unmatched knowledge of every aspect of consciousness research is an incessant inspiration to improve. It was an honor and a pleasure.

I wish to thank Larissa Albantakis, Melanie Boly, Matteo Grasso, and Jonathan Lang for their friendship, logistic support and for being always available for every question I had about IIT in its philosophical, neuroscientific, and mathematical aspects. My gratitude extends to every other student, postdoc and researchers at the Center for Sleep and Consciousness. In particular: Benjamin Baird, Leonardo Barbosa, Tom Bugnon, Anna Cattani, Matias Cavelli, Erick Chastain, Lucy Chastain, Rosario Cilento, Ogulcan Cingiler, Renzo Comolatti, Isabella De Cuntis, Colin Denis, Kort Driessen, Graham Findlay, Keiko Fuji, Andrew Haun, Jeremiah Hendren, Elsa Juan, Bjorn Erick Juel, Mariel Kalkach-Aparicio, Csaba Kozma, Sophia Loschky, Rong Mao, William Marshall, Will Mayner, Garrett Mindt, Shuntaro Sasai, Giovanna Spano, Amandine Valomon, Tom Vanasse, Ruben Verhagen, Giulietta Vigueras Borrego. I also wish to thank Adeline, Ashley, Gabriel, and Mattia for their kindness and friendship. I am glad that I had the chance to meet, work with and befriend so many wonderful people.

I wish to thank my friend Robert Chis-Ciure for our long-lasting collaboration and for countless hours of discussions about consciousness and a wide range of other topics.

I want to express my most sincere gratitude to Anna Cattani and Matteo Fecchio for their friendship and constant support, in particular during the Covid-19 pandemic.

I want to thank Matteo Capodaglio for a decade of authentic friendship despite the distance and for many valuable insights about science and life in the USA.

I want to express my recognition to Gianluca and Monica for their friendship and for teaching me some of the most valuable lessons of my life.

I also want to take this chance to thank Riccardo for being such a great friend over the past 25 years, and because through him I met the other member of "la vecchia guardia": Bruno, Fabio, Giorgio, and Marcello, whose friendship and support made all these years so much funnier.

I want to express my gratitude for their incredible support over the years and for sharing some of the most significant moments of our lives: Cristiana, Lisa, Thanos. And of course, Cristoforo, and Aléxandros for the joy of sharing their coming in the world.

Finally, I am thankful for having my parents Monica and Philippe, my sister Cécile and my late grandparents Carmen, Domenico, Jaqueline and Yvan. Nothing of what I did in my life, including this thesis, could have been possible without them.

Introduction

The story that I am about to tell has two protagonists: consciousness and the brain. Like a diarchy, they will conjointly reign over the scientific discourse which aims to answer to a simple question: "how can we fit our subjective experience into our objective description of the world?". As the tale will unfold, an unexpected plot twist will show how, we have always considered the problem from the wrong angle, and how the diarchy was a monarchy from the beginning. But for now, let introduce our protagonists.

You wake up from a night of sleep without dreams; suddenly, there is something instead of nothing. Your thoughts, emotions, memories, desires, perception of the world, dreams, ambitions and inspirations, bodily sensations, and sense of self, or, in other words, you have come into being. This is consciousness: the presence of subjective qualities, experience, what we value the most and what we understand the least.

In the duel between 'physiology and psychology,' neuroscientists always favored the former, as by nature more tractable and measurable—a task by itself of the most significant magnitude. Neurons, the collective name for over a thousand different kinds of cells, are estimated to be around 10^{11}, interconnected by approximately 10^{15} synapses in a human brain. When these numbers are combined in a connectivity matrix, they lead to many possible distinctive connectivity pathways, estimated between a googol (10^{100}) and 1 googolplex (10^{googol}). Set aside the complexity of neuroanatomy, brain activity is characterized by chaotic dynamics typical of complex systems. However, this complexity can be compressed by considering the brain as a network where its nodes (neuronal cells) are connected by edges (synapses). Hence, we can study it mathematically using graph theory. Graph theory has been considerably successful in neuroscience, leading to discovering several interesting properties of general graphs, such as small-world attributes, network communities, and rich-club networks, just to name a few (Sporns, 2011). Considering the brain from this abstract perspective presents several theoretical challenges. However, it allows us to treat the most complex object known in the universe, the brain, in a mathematical way (Ascoli, 2013).

An overwhelming body of scientific evidence links the brain with consciousness. If a boxer hits his opponent's head violently, he can knock him out. If I take substances that affect the nervous system, such as caffeine or LSD, my mind will be affected. Nevertheless, a tumor can destroy almost entirely the cerebellum, and the patient's consciousness will not be affected: not all the brain is necessary for consciousness. Even more fascinating, a vast literature on brain lesions shows that even minor damages localized in specific brain regions correlate with sensitive alterations in likewise specific conscious contents: such as the perception of colors, or human faces, or the left visual hemisphere. These qualitative aspects of consciousness just cease to exist for the affected subject, who, in return, is often not aware of their absence. It does not matter how much we discover about the brain and its relation to the conscious mind: the relation itself still seems to elude us.

In the present work, I will discuss a theoretical approach that was developed with the precise intent of clarifying how the mind and brain are related. This approach is Integrated Information Theory (IIT), developed by Giulio Tononi and his collaborators. One of the most interesting features of IIT is its phenomenology first approach and its implications for the mind-body problem. Notably, such a prominent feature did not receive much attention from critics and proponents of the theory. Therefore, one of the first goals of this thesis is to fill this gap in the literature about IIT.

In the first chapter, I introduce the mind-body problem, as the general problem of how mind and matter are related, and the hard problem, as the more specific problem of explaining how consciousness arises from matter. I then proceed to illustrate the terminology and word definitions that I will employ throughout the entire work. Finally, I present my assumptions and intuitions about the problem.

In the second chapter, I look at the mind-body problem from a historical perspective. Recently, Goff has proposed that our current neuroscientific paradigm is unable to tackle consciousness due to an error in its foundations. This error, according to Goff, is due to Galileo. According to Goff, Galileo's error is to remove qualities from physical bodies, an error due to Galileo's need to explain physical bodies in the language of the universe, mathematics, which can only deal with quantities and not qualities. I explore Goff's ideas, and I conclude that Goff is right in assuming that we need a new foundation for a science of consciousness, but for different reasons. In particular, I argue that what alienates consciousness from objective science is the primary-properties and secondary-qualities distinction. In the same chapter, I show how Descartes' dualism is a consequence of taking consciousness seriously while endorsing the primary-properties and secondary-qualities distinction.

Having shown that the hard problem is the main obstacle for any theory of consciousness and that dualism is the consequence of a wrong assumption, in the third chapter, I proceed to illustrate the axiomatic approach of IIT. Starting with consciousness and inferring the physical mechanism of consciousness, IIT twists the mind-body problem upside down and dissolves the hard problem. I discuss what can be called the 'core of the theory': its axioms, postulates, and fundamental identity. I proceed to illustrate the theory and its explanatory structure, its predictions, its explanations, and its metaphysical consequences. Finally, I address some of the most common criticism against IIT's axiomatic approach.

In the last chapter, I engage with the unfolding argument, which has been presented recently as a confutation of IIT. The unfolding argument has quickly generated an interesting debate about consciousness and falsification. In my analysis, I engage with the literature on the problem, and I show how the unfolding argument presents several criticalities. In particular, I show that the argument does not present real traits of originality; it commits to a radical definition of consciousness; it does not consider the literature from the past 100 years on the demarcation problem. Finally, I offer a response on the empirical investigation of consciousness science from the standpoint of IIT.

To conclude this introduction, I would like to discuss some of the reason that led me to believe that it was worth to pursue this project. Like many others, my fascination with consciousness started at a tender age. I loved sci-fi movies in my childhood (and I still do today), but I was confused about how inconsistently movies treated robots. Sometimes robots were portrayed like people who can feel pain and fatigue, other times as cold machines unable to feel anything. I was frustrated by the lack of "rules" to set this matter. Unwittingly, I realized the need for a principled approach to consciousness. Someone at school told me that the difference was in the soul, but it did not ease my confusion: I could not understand what it meant that humans have souls and, say, dogs do not. I could not grasp the concept of an immaterial soul; what does it mean that something exists if nobody can see it, touch it, smell it, or feel it in any way? Later on in my life, I had to deal with a sad event that sooner or later affects the life of everyone: the loss of a dear one. Nevertheless, despite the pain of the situation, we struggle to define what is exactly lost once a person is no more. Perhaps it is precisely that same bundle of feelings, thoughts, and sensations that comes into being every morning as we wake up from a dreamless sleep?

There are at least two major good reasons to write about consciousness. The first is that consciousness is both scientifically and philosophically mysterious, and human beings love to solve mysteries. Many thousands of years ago, a small group of our ancestors ventured outside their small valley in Africa, where their people have lived for generations. Within few generations, they reached almost every remote corner of

this planet, crossing forests, jungles, seas, and deserts. I like to think that they did so not only in pursuit of food, vital space, and resources but moved by that curiosity for the world and what lies beyond it that lead their descendant to invent the airplane and land on the Moon within a single lifespan. Consciousness naturally appeals to our curiosity. As much as our knowledge of the world reached an unprecedented level in human history, we have not made much progress when it comes to the most straightforward question: "How does our subjective experience fit in our objective description of the world?".

However, there is also a second reason, and it is the most important one. Consciousness is what matters for our moral decision. We use consciousness to attribute value. We do not value living people because they are alive; we value living people because they are conscious. For millennia the two things were more or less equivalent. To the death of the body corresponded the death of the brain (and therefore the mind), and vice-versa. However, with the invention of intensive care units, the death of the brain no longer implied the body's death. The price to pay for this new hope is that disorders of consciousness became a grim reality. Human bodies lying for years in hospital beds with no clarity on whether there is still someone inside or everything is gone, and only empty shell is left. On the other end of the spectrum, people perfectly conscious but trapped in the prison that their paralyzed body has become, too often misdiagnosed. A better theoretical and clinical understanding of consciousness and its underlying mechanism is imperative. The present work is a treatise primarily focused on the epistemological and ontological aspects of the scientific study of consciousness, with no ambition to provide a significant contribution for a better treatment or diagnosis of disorders of consciousness. However, I like to think that science is a massive web of intertwined practices, concepts, and problems. So, if, in the long run, my work can have even the slightest contribution to alleviating the suffering of those that are in these difficult conditions or their close ones, then I will consider my time and effort well spent.

Chapter One: Consciousness and the Mind-Body Problem

1 – Introduction

On one side are your thoughts, emotions, memories, desires, perception of the world, dreams, ambitions and inspirations, bodily sensations, and sense of self, or, in other words, *you*. On the other side are 1400 grams of cells arranged to form the most complex object in the known universe, your brain. So how does the ineffable mind arise from crude matter? This, in a nutshell, is the mind-body problem. Even without a clear definition of what is mental and what is physical, it is evident that the relation is more problematic than its relata.

The problem goes back to the origin of philosophy itself: both Plato and Aristotle discuss at length and across multiple works their ideas about the nature of the soul and its relationship with physical bodies[1]; the problem has been traditionally debated in philosophy ever since. However, it seems valid to think that the problem is even older and pre-philosophical, if for no other reason than because of the role that our intuitions play in conceptualizing the problem. Intuitively, we reckon that we are conscious. We can agree that almost certainly other people are, perhaps animals too, but hardly so 'inanimate' objects such as rocks or spoons. (Melloni et al., 2021)

We generally regard our own mind as ineffable and non-physical, while we grant that our body is physical. This clearly contrasts with the other two intuitions commonly shared: (i) a non-physical entity cannot interact with a physical entity and (ii) our mind controls our body. This first simple contradiction shows how difficult it can be to collocate the mind and body in the same framework. When we move from the terrain of intuitions to the arena of philosophical analysis, the mind-body problem presents itself as a much more complicated issue. Further, intuitions alone can no longer be the guide. In fact, if my arguments are persuasive, at the end of this journey, the reader will observe how some of the most common intuitions about consciousness and the

[1] The main dialogues in which Plato discuss the mind-body problem are *Republic*, *Phaedrus*, *Phaedo* *Cooper, J. M. (1997). Plato: Complete Works. Hackett.* . For Aristotle's discussion of soul and mind-body problem, see: *Metaphysics*, *On the Soul*, and from his minor treatises: *Sense and Sensibilia*, *On Sleep*, *On Dreams* Barnes, J. (1984). *The Complete Works of Aristotle* (Vol. I and II). Princeton University Press. .

world can be subverted. In particular, I will argue that the mind-body problem is so peculiar that its implications are foundational not only for our scientific understanding of consciousness but for science itself.

2 – What consciousness is not

Before saying what consciousness is, it might be useful to weed out some ideas that may appear reasonable but are ultimately misleading.

For example, one may be tempted to identify consciousness with self-awareness, with one's sense of self, or with the stream of thoughts made possible by language. These are all aspects of the conscious mind to which we are accustomed through our daily experiences. In fact, they are features of consciousness, not consciousness itself. Consciousness is the broad sense, the precondition for these *specific* experiences to be possible. For example, it is easy to see how, under certain conditions of depersonalization, one might lose their sense of self while being conscious. Moreover, language impairments such as aphasia leave the affected subject unable to process language but with their consciousness unaltered.

Wakefulness is not sufficient nor necessary for consciousness: a patient can be medically awake yet unconscious[2], and we are conscious while we dream. Dreaming shows us that our experiences do need a content 'out there': we can experience things while being disconnected from the environment. In fact, conceptually, there is no difference between being conscious while 'connected' to the environment (e.g., during wakefulness) and 'disconnected' (e.g., during dreams or hallucinatory experiences).

More importantly, consciousness is not behavior. We rely on inferences to navigate this ocean of complexity that is our world. Every day as I wake up, I start thinking. I become immediately aware of my thoughts as they come into being: my experience is the most direct and unfiltered thing I can imagine. I cannot see, taste, smell, hear, or touch other people's thoughts, but I know that other people belong to my species, that they look very similar to me, act as I would under any circumstances, and so on. Moreover, I have always heard people talking about their thoughts. Therefore, I conclude that most people think more or less like I do. This is an example of reliable inference. Another example could be to think that whenever someone has their eyes closed and does not move, this person is sleeping. However, it is important to notice that the person's behavior can only guide my inference. It acts like a proxy per consciousness (and not necessarily a good one, as I will illustrate later) but it is not

[2] For example, cases of unresponsive wakefulness syndrome (UWS).

consciousness itself. For example, I could be sleepwalking, or sleeping and yet experiencing countless worlds as I dwell from one dream to another. Both examples show that there is a double dissociation between consciousness and behavior: behavior is not necessary for consciousness and vice versa.

3 – Consciousness is experience

Many people consider consciousness mysterious, but we all know what consciousness is: what is present when we are awake and goes away when we sleep, returning multiple times during the night as we start dreaming (Tononi, 2004). Consciousness is everything that we experience: feelings, emotions, thoughts, perception, and so on. In other words, consciousness is experience[3]. Tastes, sounds, visual imagery, tactile sensations, thoughts, and emotions: if there is an experience, there is something, if experience is absent, there is nothing.

As noted by Finnish Philosopher and Cognitive Neuroscientist Antti Revonsuo: "In its barest essence, phenomenal consciousness constitutes an inner presence the simple presence or occurrence of experiential qualities, that is. No self is required – no representing, no intentionality, no language, no concepts – only the subphenomenal space in which phenomenal qualities may become present." (Revonsuo, 2009).

Thomas Nagel wrote a seminal article in which he wondered *what is it like to be a bat* (Nagel, 1974)[4]; such an expression became common to illustrate consciousness. Being a specific subject means feeling what that subject feels. And if my consciousness is what it is like to be me, then it follows that without my consciousness, there is no *me*. In other words, consciousness is existence[5]. Right now, I am experiencing writing these words on my computer: I see a white screen and letters appearing on it. I feel the keyboard under my fingers. At any given time, the collection of sensations (or feelings) that I am experiencing *are* my consciousness. My experiences are everything that I am, was, and will be.

---

[3] Hereon, I will use the terms 'experience' and 'consciousness' as synonyms.

[4] This expression is now commonly associated with Nagel, who popularized it in his seminal article *What is like to be a bat?* (1974). However, as Nagel himself acknowledges Nagel, T. (1986). *The View From Nowhere* (Vol. 37). Oxford University Press. , the notion of what-is-likeness was originally and independently used by Sprigge, T. (1971). Final causes. . *Proceedings of the Aristotelian Society*, *45*. while Ferrell, B. A. (1950). Experience. *Mind*, *LIX*(234), 170-198. used almost the same question, "what would it be like to be a bat," in one of his articles.

[5] This claim will be addressed further in the following chapters.

Before moving to the mind-body problem a clarification can be useful: for the entirety of the present work, I will use the terms 'consciousness', 'conscious mind', 'conscious experience', 'phenomenal mind', 'experience' and 'mind' interchangeably, as they are perfect synonyms.

4 – The Hard Problem of Consciousness

In 1884, T.H. Huxley famously said: "How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the djinn when Aladdin rubbed his lamp in the story." (Huxley, 2011). Over a century later, our knowledge of the brain as a physical and biological system has increased dramatically; yet Huxley's claim stands unchallenged. This is what is known today as the hard problem of consciousnss. One of the most intuitive ways to spell out the problem was developed from Leibniz's windmill thought experiment:

> "It must be confessed, however, that Perception, and that which depends upon it, are inexplicable by mechanical causes, that is to say, by figures and motions. Supposing that there were a machine whose structure produced thought, sensation, and perception, we could conceive of it as increased in size with the same proportions until one was able to enter into its interior, as he would into a mill. Now, ongoing into it he would find only pieces working upon one another, but never would he find anything to explain Perception" (Leibniz, 1714).

We know through introspection that the mind is rich in qualities: the redness of red, the sweetness of chocolate, the first chord of *Idomeneo,* and so on. And yet, when we peer into the human brain – not differently from Leibniz's windmill – we can find only tissues and cells. It is not a problem of finding the proper grain: whether we check brain areas, cortical columns, neurons, or even neuronal sub-components, qualities simply are not there. We could go down to the smallest atomic components of our brain and the scenario would always be the same: countless interacting mechanisms but no trace of qualities.

In more recent times, the joint endeavor of philosophy and neuroscience to explain this elusive problem has been characterized by a distinction between easy problems and the hard problem of consciousness (Chalmers, 1995; Chalmers, 1996; Chalmers, 2010). The distinction can be understood as the apparent difference in the explanations needed to account for the two sets of problems. On one hand, the easy problems are vulnerable to explanations in terms of structural configuration and

functions in physical systems: the kind of explanations traditionally employed by natural sciences, including neuroscience. Once we specify the computational and/or neural mechanism that performs the relevant function, there is nothing left to do. Hence, they are called 'easy problems'. Examples of the easy problems are those of explaining executive functions such as conscious access, reports, attention, memory, and others[6]. This type of explanation is unproblematically physical since it implies an account of the explanandum in terms of physical processes. In contrast, the hard problem refers to how we can explain *what it is like to be us* (Nagel 1974) in physical terms, i.e., the problem of explaining experience itself. Why are some physical processes accompanied by subjective qualities while others are not? Why is our brain always active but during wakefulness or dreaming we are conscious and during dreamless sleep, we are not? Why does a specific experience feel the specific way it does and not some other way? Why does mint feel like mint and not like chocolate? All these questions are at the core of the problem of consciousness (Chis-Ciure & Ellia, 2021) and no matter how hard we try, physical sciences seem inadequate to address them.

Over the past thirty years, the main focus of the neuroscience of consciousness has been the neural correlates of consciousness (or NCC) defined as the minimum neuronal mechanisms jointly sufficient for any specific conscious experience (Crick & Koch, 1990). Chalmers holds that specifying mechanisms that play some functional or causal role within a given conscious system are not sufficient to explain subjectivity or experience, but sufficient to account for any of the easy problems. However, to explain experience requires something more – even if we were to circumscribe the NCC with an extreme degree of precision, they still would not be enough to account for experience; neural areas are neural areas and experience is experience. So far, neuroscience has been able to find, at best, correlations between brain states and phenomenal states. However, these cannot be explained further; neuroscience does not give us any particular reason why, for example, activity in the amygdala correlates with fear. The hard problem is thus, apparently, resistant to the standard methods of physical sciences, which, according to Chalmers, is because it requires instead a non-reductive explanation, where consciousness itself is taken as fundamental, i.e., not explainable in simpler terms (Chalmers 1995). Clearly, the easy problems are not easy in the sense that they do not represent a significant scientific challenge. They do. But in contrast with the hard problem, they seem to be approachable within the current and familiar framework of natural sciences. Intuitively, the easy problems are problems that can be dealt with in *quantitative* terms, while the hard problem requires explaining the *qualitative* aspects of consciousness. Notably, the hard problem has little to do with the ineffability of certain experiences, i.e., the fact that they are particularly

[6] For more details on the easy problems, see Chalmers (1995, 2010).

hard or even impossible to describe, let alone to convey to another subject through language (Ellia et al., 2021).

The hard problem makes experience an explanandum in its own right. This requires no extra justification since it's something we are directly acquainted with. Furthermore, since at least the dawn of the Early Modern period, it has been widely recognized that, although there is a consistent connection between consciousness and physical matter (at least in the form of biological bodies and brains), the two seem impermeable to reconciliation and unification in a single theoretical framework. In fact, one can read Descartes' *Meditations* (Descartes, 1984) as offering an argument for the real distinction between mind and body, as I can conceive of my mind existing without my body, but I cannot coherently conceive of my body existing without itself. Thus, via the principle of the distinctness of the discernible, my mind is not identical to my body[7].

The hard problem arises due to the immediateness of phenomenal experience coupled with its stark incompatibility or incommensurability with the physical world of atoms, neurons, and bodies. Solving the hard problem is one of the greatest challenges that mankind must face in its pursuit of knowledge (Chis-Ciure & Ellia, 2021).

5 – To explain, or to explain away?

Many believe that the hard problem is ultimately an empirical problem, which will not be solved by the comfortable armchairs of metaphysicians, but requires just more empirical research to be settled (Seth, 2016). Ultimately, we know that the relationship between consciousness and the brain is tight, and certainly that the brain is more involved in consciousness than the heart or the liver, or any other part of the human body. Even though we do not know exactly how the two are related, we know that the solution must be in the brain. Therefore, the vast majority of scientists and laypeople alike are compelled, nowadays, to adopt a naturalistic attitude to the problem and endorse physicalism. Even philosophers during the last century have been progressively attracted towards physicalism, to the point that those who hold a different view are now a minority. In broad terms, naturalism is the view that metaphysics should be constrained by physical sciences. Physicalism is the view that consciousness will be ultimately explained in purely physical terms[8]. According to this view, one can claim, for example, that consciousness is identical to the brain (e.g.,

[7] See the next chapter for a detailed discussion of Descartes' meditations and the mind-body problem in the Early Modern era.

[8] For the sake of simplicity, I use materialism and physicalism as synonyms.

(Crick, 1994)), its functions (e.g., (Cohen & Dennett, 2011)) or that it is reducible to it in other ways. However, to believe that physicalism must be true because the (unclear) relation between the brain and the mind is tight is a profound misconception of the mind-body problem.

To engage with the mind-body problem means to engage with the ontology of science, which is a matter of metaphysics. To do so, one does not necessarily require new experiments. What is needed is a theoretical approach, one that may provide a hypothesis to defy the apparent logical incompatibility between experience and matter. Experiments will come in later, to check if the scientific hypothesis following the metaphysical positions can be verified. What is needed now is to rethink why our scientific methods that have been so successful in every field struggle with the one thing that we are most familiar with. While I do not desire to venture further into the debate between science and philosophy, my view on this issue can be adapted from the words of Galen[9]: the best scientist shall be also a philosopher and the best philosopher shall be also a scientist.

Since we have subjective experiences and difficulty in reconciling subjectivity within our otherwise successful scientific framework of the natural sciences, the door is opened to a dilemma that presents at least two options. One approach could be called 'conservative', i.e., taking our scientific method, its foundations, and implications as a constraint and accommodating consciousness within it. This is indeed the case for naturalistic approaches, a view that generally considers physicalism as the only plausible solution for the hard problem. The thought that there are no additional mysterious forces at play is somehow reassuring – it means that we got everything right so far, more or less. Generally, according to this view, the smallest entities postulated by physical science are considered fundamental and everything else derivative. However, a major limitation of physicalism is the lack of a clear definition of *physical*. I can certainly accept that there is no ethereal soul. But I cannot find an explanation of the mind in physical terms satisfactory if it is unclear what 'physical' means.

A different approach along this line is one that takes naturalism to its extreme: since the existence of our subjective existence is a problem for our current best scientific theories, perhaps what is wrong is not the theories but subjective experience. In other words, according to this view, consciousness does not need to be explained but explained away (Dennett, 1991). Those who support this view, in general, negate the existence of consciousness as such and consider it either an illusion (Frankish, 2016) or a misguided concept that should be eliminated from the textbook (Churchland,

[9] Galen famously said, "Quod optimus medicus sit quoque philosophus." The best physician shall be also a philosopher Singer, P. N. (2016). Galen. *The Stanford Encyclopedia of Philosophy*, *Winter 2016 Edition*. <https://plato.stanford.edu/archives/win2016/entries/galen/> .

1981; Churchland, 1986). Chalmers noted how those who hold illusionism to be true need not engage with the hard problem. However, they have a new problem called the 'intuitions problem' or the Meta Problem of Consciousness (Chalmers, 2018), which is the problem to explain why a hard problem exists if experience is merely an illusion.

Both approaches outlined above seem rather unsatisfactory. On the one hand, our experience is what we are most familiar with, the only thing that we cannot doubt[10]. On the other hand, as Leibniz remarked many centuries ago, no matter how hard we try, in the brain we can find all sorts of cells, but no qualities. So far, no scientific explanation has delivered definitive answers on the nature of consciousness and its place in the world. Maybe a different approach is needed, one that we may want to call 'radical', in the sense that requires us to rethink the foundation of our scientific method, or at least reinterpret its implications from a new perspective. Notably, this is a metaphysical operation and not an empirical one. But if one takes experience as a datum – the only one – and the physical world, including scientific observations and experiments, as an inference *within experience,* then some new considerations may be necessary.

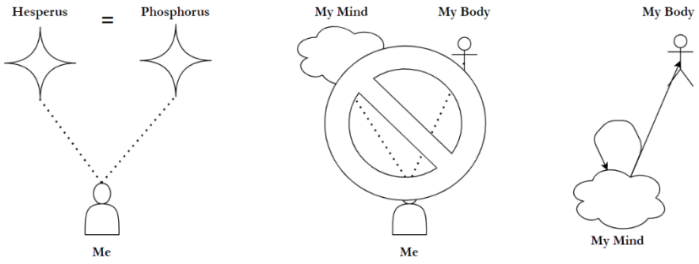
6 – A Science of Subjectivity

Before moving on, it is worth considering a further argument that shows the difficulties of standard naturalistic approaches. Consider a classic example: Hesperus (the evening star) and Phosphorus (the morning star) were considered two different celestial bodies for centuries, but we now know that they are the same planet (Venus). Whatever difference we might attribute to Hesperus and Phosphorus has to do only with the way we conceptualize them, but in reality, there is no distinction: Hesperus and Phosphorus both refer to the same thing-in-the-world. Does the same line of reasoning apply to the mind and the brain?

There is a crucial difference between consciousness and Venus (and everything else). I can easily be mistaken about Hesperus and Phosphorus being the same celestial body, but I am in a privileged position to assess their identity. As a neutral and external observer, I can conceive some experiments, register my observations, and eventually conclude that they are the same entity. When the problem is to determine whether my conscious experience is made of the same stuff as the physical world, I have already lost my neutral position. Simply enough, I cannot observe – let alone feel – anyone

[10] See next chapter.

else's experience, so I can never be in a position of neutrality with respect to the observed object as I am when it comes to Venus.



We are never 'neutral' in respect to our own experience.

As stated above, consciousness is subjective. My pain is *mine*. *I* feel it, and my friend does not. But when I think about the world, I usually assume that the world – contrary to my pain – is the same for my neighbor and everyone else. If I cease to exist, my pain ceases with me. But if I cease to exist, does the world as well? Arguably not. It is true that when I go to sleep the world ceases to exist *for me*. But when I wake up it looks like everything went on, business as usual, while I was sleeping. Commonly, we say things such as 'the world is objective' while things such as pain are subjective. What we really mean by that is ambiguous: is my pain not a real feature of the world? Is something objective somehow more real? An old thought experiment asks whether a tree that falls in a forest without anyone around to hear makes a sound[11]. Rational adults are keen to say that yes, the tree makes a sound because they have been taught that sounds are the byproduct of vibrations in the air. Children with less education may be tempted to say that no, without anyone to hear, there cannot be a sound. Kids are often wiser than educated adults.

The assumption of an objective world is what makes science possible. In fact, science is largely considered an objective description of the world, obtained through rigorous observations, manipulations, tests, and confirmations. And so far, it has worked almost perfectly. Science and the scientific method led us to a world of wonders and discoveries, from the DNA in our cells to black holes at the center of galaxies and many things that were just unthinkable a few decades ago. People on Earth and beyond can communicate in real-time independently of their location. We fly in planes in the skies, and we dive to the bottom of the oceans in submarines. We can pilot rovers on planets that we started to look at with a telescope only a few centuries ago. All these achievements are almost incredible once we realize they were all made

[11] The thought experiment is often misattributed to Berkeley. One of the closest sentences he wrote can be found in his Treatise: "The objects of sense exist only when they are perceived; the trees therefore are in the garden... no longer than while there is somebody by to perceive them." Berkeley, G. (1734). *The analyst: A discourse addressed to an infidel mathematician.* Wilkins, David R. .

possible in a few centuries and thanks to the scientific method[12]. Science has been proved so successful that anyone who doubts its foundation should be looked at with suspicion. It almost feels like, piece by piece, science is solving the gargantuan puzzle of reality. Yet, this almost perfect and objective description of the world falls short when it has to account for what is most intuitively true for us: our subjective experience.

Whether we are lay people or scientists, we do not just experience the world; we experience it from our subjective point of view. Each one of us has a unique and specific point of view on reality. How can we reconcile this with the objective – and successful – description of the world provided by science? It simply seems that there is no place for subjectivity in our scientific picture. Physics, chemistry, biology, and even the social sciences tell us about a world seen in its entirety notwithstanding any particular point of view, a third-person description of reality that philosopher Thomas Nagel called 'the view from nowhere' (Nagel, 1986). Ultimately, any scientific theory aiming to provide a complete description of the world has a major challenge: to provide a principled answer to this fundamental question:

How is it possible to fit our subjective experience into our objective descriptions of the world?

The implications of this question are far broader than they may appear. In fact, the standard approach in neuroscience is to attempt to derive the subjective (consciousness) from the objective (the brain). While the connection, currently, is still unclear, it must be there. However, two separate considerations can be helpful here. First, as noted above, our subjective experience is most directly available to us. We are directly acquainted with our own experience. The 'external' and objective world by contrast is only inferred from our conscious experience. It is an extremely good and reliable inference, but indeed only an inference. Moreover, every human activity – including science – starts with a conscious subject. It does not matter what sophisticated tools or mathematical models we employ to investigate reality; ultimately, we know the objective world only because of and within our subjective experience. Therefore, a more promising endeavor could be to revert the order, starting with what is most known to us (our subjective experience) and moving to what is less known to us (the external world)[13].

Put in these terms, it does not seem too excessive to claim the mind-body problem is unique and different from any other philosophical and scientific problem, not only because it problematizes the relation between the physical world and the ineffable mind but also because of its foundational implications.

[12] I intend science in the broadest acceptation of the expression.
[13] This line of reasoning will be considered more in depth in the third chapter.

7 – Conclusion

The hard problem raises many questions about experience and its place in nature. Why is there consciousness in the first place? How does the ineffable mind emerge from matter? Why are some neural processes accompanied by subjective feelings while others are not? More in general, why are certain physical systems under certain conditions conscious while others are not? Why do qualities feel the way they do and not a different way? In other words, how can we fit our subjective experience into our objective description of the world? Such a foundational question cannot be decided with data alone but requires a highly theoretical approach.

In the rest of this work, I will illustrate what I believe is the most promising approach to this problem. This approach not only offers a solution, but one that, in principle, can be tested *empirically*. The approach I will discuss for the rest of this work is the Integrated Information Theory (henceforth, IIT), developed by Giulio Tononi and his collaborators over the past two decades (Oizumi et al., 2014; Tononi, 2015; Tononi et al., 2016).

Bibliography

Albantakis, L. (2021). My conscious(ness) biases. *Conscious(ness) Realist*. https://www.consciousnessrealist.com/consciousness-biases/

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, *21*(5), 459. https://www.mdpi.com/1099-4300/21/5/459

Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy*, *23*(3), 362. https://www.mdpi.com/1099-4300/23/3/362

Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), 18803. https://doi.org/10.1038/s41598-020-75943-4

Barnes, J. (1984). *The Complete Works of Aristotle* (Vol. I and II). Princeton University Press.

Berkeley, G. (1734). *The analyst: A discourse addressed to an infidel mathematician*. Wilkins, David R.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200-219.

Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9-10), 6-61.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press.

Chis-Ciure, R., & Ellia, F. (2021). Facing up to the Hard Problem of Consciousness as an Integrated Information Theorist. *Foundations of Science*. https://doi.org/10.1007/s10699-020-09724-7

Churchland, P. M. (1981). Eliminative Materialism and Propositional Attitudes. *Journal of Philosophy*, *78*(2). https://doi.org/10.5840/jphil198178268

Churchland, P. S. (1986). *Neurophilosophy: Toward A Unified Science of the Mind-Brain* (Vol. 97). MIT Press.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8). https://doi.org/10.1016/j.tics.2011.06.008

Cooper, J. M. (1997). *Plato: Complete Works*. Hackett.

Crick, F. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul* (Vol. 37). Scribners.

Crick, F., & Koch, C. (1990). Towards a Neurobiological Theory of Consciousness. *Seminars in Neuroscience*, *2*, 263-275.

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Descartes, R. (1984). *Meditations on First Philosophy*. Caravan Books.

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, J., Haun, A., Albantakis, L., Boly, M., & Tononi, G. (2021). Consciousness is a structure, not a function. *In preparation*.

Ferrell, B. A. (1950). Experience. *Mind*, *LIX*(234), 170-198.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11-12), 11-39.
Gomez, J. D., Mayner, W. G. P., Beheler-Amass, M., Tononi, G., & Albantakis, L. (2021). Computing Integrated Information (Φ) in Discrete Dynamical Systems with Multi-Valued Elements. *Entropy*, *23*(1), 6. https://www.mdpi.com/1099-4300/23/1/6
Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, *21*(12). https://doi.org/10.3390/e21121160
Huxley, T. H. (2011). *Collected Essays*. Cambridge University Press.
Leibniz, G. W. (1714). *Monadology*.
Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: A toolbox for integrated information theory. *PLOS Computational Biology*, *14*(7), e1006343. https://doi.org/10.1371/journal.pcbi.1006343
Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*(October), 435-450.
Nagel, T. (1986). *The View From Nowhere* (Vol. 37). Oxford University Press.
Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10*(5). https://doi.org/10.1371/journal.pcbi.1003588
Revonsuo, A. (2009). *Consciousness: The Science of Subjectivity*. Psychology Press.
Seth, A. (2016). The real problem. *Aeon*.
Singer, P. N. (2016). Galen. *The Stanford Encyclopedia of Philosophy*, *Winter 2016 Edition*. <https://plato.stanford.edu/archives/win2016/entries/galen/>
Sprigge, T. (1971). Final causes. . *Proceedings of the Aristotelian Society*, *45*.
Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42
Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. https://doi.org/10.4249/scholarpedia.4164
Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450-461. https://doi.org/10.1038/nrn.2016.44

Chapter Two: Quantities and Qualities

1 – Introduction

As illustrated in the previous chapter, the hard problem of consciousness seems to resist the standard methods of physical sciences. Then perhaps a useful strategy could be to go back to the foundation of modern science and try to understand why subjective experience does not fit our objective worldview. Clearly such a task is of the greatest magnitude and perhaps not even possible in its entirety. Therefore, I will reduce the scope of my inquiry and focus exclusively on two seminal works: *The Assayer* by Galileo Galilei and the *Meditations* by René Descartes.

I will argue that it is not by coincidence that the mind-body problem became such a crucial issue in the Early Modern period, and that the work of these two most influential thinkers of the Western tradition played a central role. I will argue instead that they laid the foundations of our way to think about the physical and the mental and, therefore, about the mind-body problem.

2 – Galileo's Error

Galileo Galilei is commonly regarded as one of the fathers of modern science. Before Galileo, the predominant view on the philosophy of nature (i.e., the discipline that today we call physics) was based upon the theories and the ideas of Aristotle. Part of the philosophical background on his new method can be found in a little text in the form of a letter to Pope Urban VIII: *Il Saggiatore*, or *The Assayer*.

Galileo wrote *The Assaye*r in 1623, a treatise about comets, defined as 'the greatest polemic ever written in physical science' (Drake, 1957). Today, *The Assayer* is considered a fundamental piece in the evolution of scientific thought. The polemic at the center of this book started five years prior to its publication when three comets appeared in the sky of Europe and became a point of debates among philosophers and people of science. Among them, Orazio Grassi published *De tribus cometis anni MDCXVIII disputatio astronomica,* where he argued that comets were located beyond the moon. Grassi was a follower of Tycho Brahe's geocentrism, which adhered to Aristotelian science; Grassi's arguments were based on pure logic rather than empirical observations. Galileo was so angered by this that apparently his copy of Grassi's work was annotated with insults of all kinds, such as *pezzo d'asinaccio* (piece of a bad donkey) and *balordone* (bumbling idiot) (De Santillana, 1955). Galileo's reply came the same year through the pen of one of his students, Mario Guiducci, who wrote *Discorso delle Comete*, in which he defended Copernicus' view and argued (erroneously) that comets were not 'real' objects, but mere illusions originating within Earth's atmosphere and therefore sublunar. Grassi, who did not mean to yield, wrote a second reply, the *Libra astronomica ac philosophica*, this time under the pseudonym of Lotario Sarsi Sigensano[14]. The polemic went on and Galileo bit the bullet pretending to believe that Sarsi and Grassi were two different people (although he did not miss the chance to refer to him as the 'unheard Sarsi'), and he wrote another reply: *Il Saggiatore* or *The Assayer*. The contempt Galileo held for his opponent and the irony that characterizes the whole work is evident from the title[15]: *libra* is a kind of scale used to weigh big items, a steelyard balance, but not a very precise one: according to Galileo, this is the scale that Sarsi used to weigh his ideas[16]. On the contrary, the *saggiatore* is a precision scale, the kind used by goldsmiths, and in the introduction, Galileo promised that he would use the same care that the goldsmith uses to weigh gold to weigh his ideas. The book was an immediate success, and the fact that it was written in Italian contributed to Galileo's fame in his home country; however, for the same reason, the book was less known abroad.

[14] Notice how Grassi's chosen pseudonym is the anagram of his Latin name: Horatius Grassius Salonensis.
[15] The meaning of the title is debated in the literature: Bianchi (2014) argues that the correct acceptation of the title Il Saggiatore is 'the man who assays'.
[16] It seems to be a clever wordplay with the title of Grassi/Sarsi's book: *Libra astronomica ac philosophica.*

The Assayer was first and foremost an astronomy treatise, whose primary goal was to clarify what is a physical body, such as a comet, and hence set the dispute. By doing so, Galileo precipitated the twilight of the old Aristotelian science and the dawn of the new scientific worldview. His argument presents two critical elements: on one hand, the superiority of a method bounded by empirical observation rather than purely logical arguments; on the other, the introduction of a distinction between how the world really is and how we see it, i.e., the distinction between primary properties and secondary qualities that had an enormous influence on subsequential philosophy. Locke, Boyle, Hobbes, and Descartes also used a similar distinction, but with different names. According to Hume, the distinction between primary properties and secondary qualities is a key element of modern philosophy:

> The fundamental principle of that philosophy is the opinion concerning colours, sounds, tastes, smells, heat and cold; which it asserts to be nothing but impressions in the mind, deriv'd from the operation of external objects, and without any resemblance to the qualities of the objects. […] This principle being once admitted, all the other doctrines of that philosophy seem to follow by an easy consequence. For upon the removal of sounds, colours, heat, cold, and other sensible qualities, from the rank of continu'd independent existences we are reduc'd merely to what are called primary qualities as the only real ones, of which we have any adequate notion (Hume, 1736/1992: 226).

Such a distinction traces a line between what depends upon the observer and what does not, i.e., what we may want to call mind-independent entities and what one may even be tempted to appeal as objective. In Galileo's words:

> Now I say that whenever I conceive any material or corporeal substance, I immediately feel the need to think of it as bounded, and as having this or that shape; as being large or small in relation to other things, and some specific place at any given time; as being in motion or at rest; as touching or not touching some other body; and as being one in number, or few, or many. From these conditions, I cannot separate such a substance by any stretch of my imagination. But that it must be white or red, bitter or sweet, noisy or silent, and of sweet or foul odor, my mind does not feel compelled to bring in as a necessary accompaniment. Without the senses as our guides, reason or imagination unaided would probably never arrive at qualities like these. Hence, I think that tastes, odors, colors, and so on are no more than mere names so far as the object in which we place them is concerned, and that they reside only in consciousness. Hence if the living creature were removed, all these qualities would be wiped away and annihilated. But since we have imposed upon them special names, distinct from those of the other and real qualities mentioned previously, we wish to believe that they really exist as actually different from those (Galilei, 1957: 274).

The idea is simple: if our imagination imposes qualities upon physical bodies then our imagination can strip them off again. The result is a concise and clear definition of what a physical body is: a bundle of primary properties (i.e., 'primary accidents' - or *primi e reali accidenti* in Galileo's terminology). In Galileo's view, primary properties are objective (i.e., mind-independent) and relational[17]. Most importantly, and contrary to the Aristotelian doctrine, qualities do not belong inherently to physical bodies but are relegated within the observer. In other words, the secondary qualities (i.e., 'affections' or *diverse affezioni*) are mind-dependent. Physical bodies are no longer conjunction of matter and form as in Aristotelian 'corporeal substance'. This was the real break between Galileo and Aristotelianism: a new way to think about the world, a world that can be observed and described independently by the person observing and describing it. The most cited quote of the entire book is Galileo's poetic proclamation of mathematics as the language of the book of nature:

> Nature is written in this grand book, the universe, which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometrical figures, without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth (Galilei, 1957).

The dispute on the nature of comets can finally be settled, if, under this new definition of a physical body, there is nothing left once all qualities are stripped away, then Sarsi is wrong, and the three comets are mere illusions. As we now know, in this specific matter, Galileo was wrong as comets are indeed physical bodies. But we know this because our skills and technologies to observe and understand the physical world have improved with time, and Galileo played a central role in laying the foundations for this successful scientific method.

3 – Galileo's Error's Error

Galileo's *The Assayer* is a milestone of Western scientific thought, and it has serious consequences for the way we conceive of science in general, and, therefore, the way we conceive of the science of consciousness in particular. In his 2019 book, *Galileo's*

[17] Primary properties are shape, size, location, contiguity, and motion, always defined in terms of other bodies: such as bigger than, touching or not touching another body, and number. Galileo is explicit on the fact that numbers are not absolute but relative (Galilei, 1957: 241).

Error, philosopher Phillip Goff claims that the problem of consciousness that we face today in neuroscience – the mind-body problem – is traceable all the way back to the origin of modern science. More precisely, according to Goff, the present mind-body problem is due to a conceptual mistake Galileo made in *The Assayer* (Goff, 2019).

Today, we learn something new every day about the brain and how it works, but when it comes to its relationship with the ineffable mind, we are still in the dark. Many thinkers[18] believe that this is only a problem of quantity: we just need to do more research – once we know enough about the brain, everything will be clear about consciousness. According to them, the problem is not with the methods of physical science that have been proved successful for centuries over an incredible range of natural phenomena; if we keep doing what we are doing, sooner or later the full picture will look clearer, and the mystery may even disappear altogether. In short, those researchers do not think that we need to do something qualitatively different in terms of a research program. Goff disagrees. According to him, there is not simply a missing piece of the puzzle[19]; he argues that it is obvious that consciousness is eluding natural sciences: this is precisely how natural sciences were 'designed' by Galileo. Hence, Goff claims we need a new foundation for the science of consciousness (Goff, 2019). He argues for an elegant solution, a form of panpsychism that can reconcile physical and phenomenal properties by grounding the former on the latter[20].

According to Goff, Galileo's error was to relegate consciousness outside the domain of science. He argues that Galileo was forced to do so, as he believed that mathematics cannot deal with qualities (but only with quantities) and he was seeking to found natural sciences over mathematics:

> What is so special about the characteristics of size, shape, location, and movement? The crucial point is that these characteristics can be captured in mathematics. Galileo did not believe that you could convey in mathematical language the yellow color or the sour taste of the lemon, but he realized that you could use a geometrical description to convey its size and shape. And it is possible in principle to construct a mathematical model to describe the motion of, and the relationships between, the lemon's atoms and subatomic parts (Goff, 2019: 22).

If secondary qualities cannot be expressed in mathematical terms, then this is why they cannot find a place within the Galilean framework. Therefore, there is no mystery here: we are unable to fit our subjective experience into our objective description of

[18] See for example: Graziano, 2019; Frankish, 2017; Seth, 2016; Dennett, 1990; Churchland, 1985.
[19] To be more precise, Goff is well aware that our current knowledge of the brain is far from complete; however, he does not think that this is the point.
[20] An exhaustive presentation of Goff's favored solution to the mind-body problem can be found in his two books (Goff, 2017; Goff, 2019).

the world precisely because subjectivity is inherently qualitative and our conceptual framework to think about the world does not accommodate qualities *ab origine*.

> Galileo's philosophy of nature has also bequeathed us deep difficulties. So long as we follow Galileo in thinking (A) that natural science is essentially quantitative and (B) that the qualitative cannot be explained in terms of the quantitative, then consciousness, as an essentially qualitative phenomenon, will be forever locked out of the arena of scientific understanding (Goff, 2019: 26).

In short, Goff's premise for a new foundation of consciousness science relies on Galileo's supposed original sin: 'The problem of consciousness began when Galileo decided that science was not in the business of dealing with consciousness' (Goff, 2019: 27).

There is only one problem with this thesis: Galileo does not claim that mathematics cannot express qualities. Furthermore, Galileo never argued that the use of mathematics to 'read' the book of Nature (i.e., the world) is made possible by the fact that qualities are not part of it. On the contrary, his whole thesis is that mathematics captures primary properties well, and primary properties are what constitute a physical body. As Koestler (2014) remarks, many things commonly accepted about Galileo are not true[21]. For the rest of this section, in the same polemic spirit of Galileo*,* I will argue against Goff's interpretation of *The Assayer* in his *Galileo's Error* and present my alternative reading of this issue.

First, a clarification on the mathematical nature of primary properties is due. Recall what are primary properties according to Galileo: shape, size, location, and contiguity, all of which can be considered *geometrical* properties. Then there is the number, which is an *arithmetical* property. Finally, there is motion. In Galileo, motion is neither geometrical nor arithmetical. Moreover, motion is not a mathematical property at all. It can be expressed in mathematical terms but is not purely mathematical per se (Buyse, 2015). By appreciating this fact, we notice that primary properties are not *essentially* mathematical, but they can be expressed in a mathematical way. It is a subtle difference, but highly relevant in this context. In fact, for the same reason, we should not be eager to discard primary qualities from science if they can be expressed in a mathematical way even though they are not essentially mathematical.

But the real point of disagreement that I have with Goff concerns a specific passage in Galileo's original text: Galileo supposedly claimed that qualities are to be found within consciousness. In the original Italian text, nothing seems to suggest so. The

[21] Galileo did not invent the telescope, nor the microscope, nor the pendulum clock. He did not discover the sunspot, nor did he throw weights from the leaning tower of Pisa. He was not tortured by the inquisition and almost certainly did not say *'eppur si muove!'* after his conviction (Koestler, 2014).

problem, I argue, is that Goff follows Stillman Drake's version of *The Assayer*, which was first published in 1957 and remained for over half a century the only English translation of *The Assayer* (Buyse, 2015). Consider the following text:

> Hence, I think that tastes, odors, colors, and so on are no more than mere names so far as the object in which we place them is concerned, and that they reside only in consciousness. Hence if the living creature were removed, all these qualities would be wiped away and annihilated (Galilei, 1957: 274).

However, this is not what the original Italian text says:

> Per lo che vo io pensando che questi sapori, odori, colori, etc., per la parte del suggetto nel quale ci par che riseggano, non sieno altro che puri nomi, ma tengano solamente lor residenza nel corpo sensitivo, sì che rimosso l'animale, sieno levate ed annichilate tutte queste qualità; (Galilei, 2017/1623).

Galileo never claimed that qualities are within consciousness and that consciousness was outside the domain of science. In fact, he wrote that they are within 'the sensible body' (*corpo sensitivo*) i.e., in the body of the observer (Buyse, 2015). Goff himself acknowledges that contrary to Descartes, Galileo followed Aristotle in the conception of the soul as essentially embodied (Goff, 2019). The mind-body distinction that Goff is superimposing following Drake was not present in the original text. Recently[22], the same point was raised by neuroscientist Christof Koch in his review of Goff's book (Koch, 2020). Koch's interpretation, much more in line with Galileo's writing, is that the Italian astronomer assumed a pragmatic stance to apply two new tools that recently appeared in the scientific debate about the physical universe – the telescope and mathematical physics – but he was not really interested in what, today, we call consciousness. Hence, Koch concludes, 'To me, it seems that Goff is retrofitting his ideas of the mind onto Galileo' (Koch, 2020). This idea is supported by the fact that, as I remarked above, *The Assayer* is first and foremost a treatise on astronomy. While Galileo was certainly interested in providing an alternative way to the Aristotelian doctrine to think about the universe, his main urge was to settle the controversy about the three comets that appeared in Europe five years earlier.

4 – Galileo's real error

[22] This chapter was originally written before Koch's review of *Galileo's Error* was made public and later edited to include its important insights.

So, what was Galileo's error, if any? Once again, I am inclined to think that it is not fair to frame the thought of an Early Modern era thinker in terms of our contemporary debate. Rather than attribute a specific error to Galileo, I will point out what, within *The Assayer*, collides with our way of thinking about consciousness. Perhaps the best way to start is with one simple question: when Galileo proposed a distinction between primary properties and secondary qualities, what was he actually doing? The classic answer would be that he wanted to provide an objective method to do science in contrast to that of Aristotle: no more qualities but just *facts*. But what are facts, exactly? And how can we know about them? For Galileo, the answer is easy: our mind imposes secondary qualities over bodies, and therefore, our imagination can strip them off again. But why, for example, should the redness of the table be an artifact of my mind while its shape is not? After all, I know about both in the same way: by experiencing them. The level of abstractness that we employ in our thoughts does not matter; everything we know about, we know *within* our conscious experience.

There is a hidden assumption in his way of thinking: realism, i.e., the idea that there is a world beyond our senses, and we interact with it. John Locke (1632 – 1704) – who popularized the distinction between primary properties and secondary qualities among philosophers[23] – argued that to embrace such distinction we need to commit to a particular kind of realism: critical realism. Contrary to standard naïve realism, i.e., the idea that the world is exactly as we perceive it, Locke argues for a critical realist position. Critical realism implies that the world as we experience it is not an exact copy, but a 'filtered' one. Our senses do not simply capture the world as it is but alter it slightly. While perception in a naïve realist account is purely a passive feature, a critical realist account implies that our perceptions are not identical to the world but correspond to it[24].

There is no way to tell if Galileo ever thought about the problem in these terms, but under this light, the distinction between primary properties and secondary qualities is justified by what is less variant across different observers. Whoever is going to measure the table will say that it is 6 ft long. But different people could describe its color as maroon, vermillion, sangria, or any other shades of red. Only with more precise methods can we determine a stable description of its color. For example, we can determine what is the frequency of the light reflected by the table, hence an 'objective' description of its color. Primary properties express (mathematically) qualitative aspects of reality; they are just less varied across different subjects. Therefore, they are easier to render in mathematical terms or convey verbally. But primary properties, as everything else, are experienced *within* consciousness.

[23] Recall that The Assayer was written in Italian and not in Latin, therefore it is disputed if Boyle, Locke, and other Galileo's contemporaries had directly access to it (Anstey, 2000: 25).
[24] For a general introduction on Locke see: (Uzgalis, 2020). For a better characterization of its thought on perception and reality see Woolhouse (1971, 1983, 1988).

Therefore, they cannot be used to justify in principle an assumption of realism, whether critical or naïve.

The problem then is not within Galileo's work necessarily but with the use that people made of his method after him. Galileo, without saying much about the nature of the mind, suggested a world that could be studied objectively by distinguishing between primary properties and secondary qualities, without realizing that even primary properties appear the way they do only because they are experienced by a subject. Even things such as space and motion have qualitative aspects (Haun and Tononi, 2019; Lotze 1884; James, 1874).

In other words, Galileo snuck into modern science the idea that any observer, by abstracting from her experience, can reach an ultimate and objective reality prescinding from her point of view. But the observer is always bound by her experience and cannot go beyond it. Therefore, what we call 'objective' is merely shared across many different subjects or intersubjectively.

5 – Descartes' Doubt

The scientific revolution that Galileo contributed to set in motion had a revolutionary impact not only within the scientific debate but also over the philosophical one. René Descartes (1596 – 1650), known also as Renatus Cartesius[25], is often considered the father of modern philosophy. However, during his life, he was first and foremost a mathematician, with his main contribution to mathematics amounting to the tools and techniques that made algebraic geometry possible. As a philosopher of nature – what we would call a scientist today – his contributions range from hypotheses about the formation of planets to a naturalistic account of rainbows and being the co-framer of the sine law of refraction. We remember him as a metaphysician, for providing a portrait of a physical world ontologically dissociated from the ethereal minds who populate it (Hatfield, 2018).

Contemporary philosophers and neuroscientists are fast to discard Descartes' dualism as an anti-scientific option:

> Since it is widely granted these days that dualism is not a serious view to contend with, but rather a cliff over which to push one's opponents (Dennett, 1978).

[25] The commonly used adjective 'Cartesian' derives from his Latin name.

Books have been written about the influence of Descartes' works on modern science (e.g., Damasio, 1994). However, in their criticism, they usually go after the implications of this position rather than exploring its origin[26]. In this section, I will offer a guided reading of Descartes' *Meditations on First Philosophy* focusing on the arguments for the mind-body distinction as they are exposed in meditations one, two, and six. I will argue that Descartes' arguments are valid; however, they follow from a crucially wrong premise, the primary properties-secondary qualities distinction that Galileo introduced earlier. Therefore, I do not challenge the validity of the argument but its soundness. In other words, I challenge that the mind-body distinction is real (and not only conceptual) and therefore that we should not believe in dualism, i.e., the metaphysical thesis that mind and body are made of two ontologically different substances.

6 – The First Meditation

In his *Meditations on First Philosophy*, Descartes set the goal to investigate the foundations of all his knowledge and discuss the nature of his mind, the physical world, and God. To do so, the French philosopher decided to think about those things as if no one ever wrote about them, to be free from any constraint and let his philosophical acumen wander freely.

The first meditation begins with explicating the goal of the entire project: to attain true knowledge, one has to demolish all his false beliefs, so that one can be free to rebuild on new foundations. Moreover, false beliefs are not the only misleading and uncertain entities. Almost like in a trial, all Descartes needs is reasonable doubt. In his words: 'So, for the purpose of rejecting all my opinions, it will be enough if I find in each of them at least some reason for doubt' (Descartes, 1641/1996:12). If some beliefs are doubtful, how can they be trusted to build solid foundations? Descartes argues that he needs to find the basic principles of knowledge:

> Once the foundations of a building are undermined, anything built on them collapses of its own accord; so I will go straight for the basic principles on which all my former beliefs rested (Descartes, 1641/1996: 12).

We know that some beliefs we hold (or held in the past) can be deceptive; in particular, our senses trick us constantly into believing things that are different from reality, and therefore, the beliefs that we form upon our sensory experience are the most

[26] See for example Foster (1993) for a reply to Dennett's criticism of Descartes.

deceptive. For example, big objects appear to be small from a distance and so on. However, such everyday examples of how our senses illude us are nothing compared to other kinds of deceptions. Descartes considers the case of a man gone mad due to brain injury[27] who lives in a world of his own made by hallucinations. But he then realizes that we all live in worlds of our own, without the need to become mad, every night when we fall asleep and dream. He recalls that his dreams, no matter how strange, seem always truthful to him while he is dreaming. Therefore, there is no way to tell if Descartes is wakeful and thinking about his dreams or if he is dreaming right now. Perhaps he is not even Descartes, and the apparently familiar world he lives in is being conjured by his dreaming mind. However, the French philosopher concludes that the thing that he has experienced during his dreams must refer to things that exist, as much as a painter cannot conjure unreal images but only copy reality or create compositions of real things.

That said, Descartes moves on into describing the properties of these real things. After reading *The Assayer,* they should look familiar:

> This class appears to include corporeal nature in general, and its extension; the shape of extended things; the quantity, or size and number of these things; the place in which they may exist, the time through which they may endure, and so on. (Descartes, 1641/1996: 14)

He is describing precisely the primary qualities described by Galileo! This is very important because it shows what a physical object is for Descartes: exactly what it was for Galileo. With all the consequences noted above.

7 – The Second Meditation

The second meditation starts where the first one stops, with the introduction of radical skepticism i.e., the philosophical doctrine that asserts that nothing can be known.

> I will suppose then that everything I see is spurious. I will believe that my memory tells me lies, and that none of the things that it reports ever happened. I have no senses. Body, shape, extension, movement and place are chimeras. So, what remains true? Perhaps just the one fact that nothing is certain (Descartes, 1641/1996: 16).

[27] More precisely, Descartes follows the humoral theory and consider madness those that are driven insane by the vapors of melancholia (Descartes, 1641/1996: 13).

Descartes introduces skepticism only to disprove it, which he does through the famous philosophical argument: *cogito ergo sum*[28]. In fact, the French philosopher argues that everything is open to doubt (because of the aforementioned scenario) and yet one thing is certain and indubitable: the fact that he is doubting. But in order to be doubting, there must be something that is to be doubted, or in other words, has to exist:

> Yet apart from everything I have just listed, how do I know that there is not something else which does not allow even the slightest occasion for doubt? Is there not God, or whatever I may call him, who puts into me the thoughts that I am now having? But why do I think this, since I myself may perhaps be the author of these thoughts? In that case am not I, at least, something? But I have just said that I have no senses and no body. This is the sticking point: what follows from this? Am I not so bound up with a body and with senses that I cannot exist without them? But I have convinced myself that there is absolutely nothing in the world, no sky, no earth, no minds, no bodies. Does it now follow that I too do not exist? No: if I convinced myself of something then I certainly existed. (Descartes, 1641/1996: 16-17)

Even if *there is* a deceiving devil, 'let him deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something' (Descartes, 1641/1996). I think therefore I am.

This also leads Descartes to rethink what he is, or more in general, what human beings are. Not rational animals, as Aristotle posited almost 2000 years earlier, but thinking things: 'I am, then, in the strict sense only a thing that thinks' (Descartes, 1641/1996: 18). The implications are striking – by describing himself essentially as a thinking thing, he means that all his other characteristics are accidental, i.e., can be stripped away without altering his essence. But without our minds, we simply do not exist.

In other words, consciousness equals existence. We exist because we are conscious. Without our consciousness, we do not lack *something*; we lack *everything*: We do not exist as subjects. This characterization of existence is crucial:

> At last I have discovered it – thought; this alone is inseparable from me. I am, I exist – that is certain. But for how long? For as long as I am thinking. For it could be that were I totally to cease from thinking, I should totally cease to exist (Descartes, 1641/1996: 18).

Now the challenge for Descartes is to understand whether his mind and body are two different things. To do so, he posits what can be understood in modern terms as a conceivability argument. Clearly it is possible to think about the mind without

[28] I am grateful to Erick Chastain for showing me that a similar argument, *si fallor sum* (if I am mistake, I exist) was made centuries earlier by Saint Augustine of Hippo in his *City of God.* (Trape et al. 1965/2010). The same remark was also made by Koch (2019).

considering the body, as much as it is possible to do the contrary. Remember that a body for him amounts to a bundle of primary properties:

> By a body I understand whatever has a determinable shape and a definable location and can occupy a space in such a way as to exclude any other body; it can be perceived by touch, sight hearing, taste or smell, and can be moved in various ways, not by itself but by whatever comes into contact with it (Descartes, 1641/1996: 17).

But Descartes here is looking for something more subtle: he argues that it is possible to conceive a mind without a body but is not possible to conceive a body without itself. This seems to threaten the conceivability of mind and body as the same entity. If Descartes can push this argument forward, then he will have an argumentative proof of dualism.

8 – The Sixth Meditation

The subtitle of the sixth meditation exhaustively describes its content: 'concerning the existence of material things and the real distinction between mind and body'. The aim of this meditation is to argue for a *distinctio realis*[29] between the mind and the body by an appeal to a difference in their essential properties.

As shown in the second meditation, it is possible to conceive a mind without a body. But what Descartes is after is not merely a conceptual difference between the two, but a *real* one. Consider the following example: I can think of René Descartes as a philosopher but also as a mathematician. Let's call them RDP and RDM, respectively. I can even argue that RDM was more accomplished in his career than RDP, or that students still read RDP's works while the same is not true for RDM. However, all these differences are purely conceptual. They are not real; when we look into the world, we realize that RDP and RDM are necessarily not distinct, and they are *in essence* the same thing: the philosopher and mathematician René Descartes. Therefore, the mere conceivability of a disembodied soul is not enough to argue for its ontological separation from the body – a divergence in the essential natures of the two is required. Descartes argues that the mind-body divergence subsists based on him noticing how the body is essentially extended while the mind is not[30].

[29] A real distinction and not a merely conceptual one.
[30] A different conceivability argument is used by Chalmers to show the validity of the hard problem. See chapter five for a discussion.

His argument implies accepting that material things exist with a high probability, if for no other reason than because they are the subject matter of pure mathematics[31]. To support this claim, the French philosopher offers an argument that shows the differences between imagination and pure understanding:

> When I imagine a triangle, for example, I do not merely understand that it is a figure bounded by three lines, but at the same time I also see the three lines with my mind's eye as if they were present before me; and this is what I call imagining. But if I want to think of a chiliagon, although I understand that it is a figure consisting of a thousands sides just as well as I understand the triangle to be a three-sided figure, I do not in the same way imagine the thousand sides or see them as if they were present before me. It is true that since I am in the habit of imagining something whenever I think of a corporeal thing, I may construct in my mind a confused representation of some figure; but it is clear that this is not a chiliagon. For it differs in no way from the representation I should form if I were thinking of a myriagon, or any figure with very many sides. Moreover, such a representation is useless for recognizing the properties which distinguish a chiliagon from other polygons. But suppose I am dealing with a pentagon: I can of course understand the figure of a pentagon, just as I can the figure of a chiliagon, without the help of the imagination; but I can also imagine a pentagon, by applying my mind's eye to its five sides and the area contained within them. And in doing this I notice quite clearly that imagination requires a peculiar effort of mind which is not required for understanding; this additional effort of mind clearly shows the difference between imagination and pure understanding (Descartes, 1641/1996: 51).

The same argument also serves the purpose to show that, contrary to understanding, imagination is not a constituent of the mind's essence: a mind without imagination may be plain, but a mind without understanding is meaningless; it loses its essence. Neither imagination nor perception could exist without the mind that contains them. So, the mind itself is the most fundamental entity, and its essence is thought.

What about physical objects? Throughout previous meditations, Descartes argues that God is no deceiver. Hence, the world might be different from what it appears to him but not radically so. He may not be able to grasp the essence of things and his experience can be obscure and confused but the things he experiences must refer to a genuine entity[32]. Among those entities, the one he is the most confident about is his own body. Our bodily sensations are particularly strong, and it is undeniable that we

[31] This may look confusing, as one of the premises is that nothing about the world is certain. In the First Meditation, he did not allow himself to take anything for granted given that he was looking for a foundation of his method of knowledge. Now that the foundation has been achieved, through the *cogito ergo sum* argument, Descartes can accept that certain things most probably exist, even without certainty.

[32] Here the parallel between Locke and Descartes is particularly strong (Woolhouse, 1971).

experience our bodies differently from other things. Consider, for example, the feeling of thirst or hunger and how pervasive they can be in our own experiences. However, not even these sensations are experienced clearly and distinctively as they should be according to understanding but are confused and obscure: the mind has a limited capacity to interpret what happens to the body, so, Descartes argues that the body is not part of the mind.

The conclusion of the argument comes with the characterization of extendedness as the essential property of physical entities. Bodies are divisible, in the sense that they can be partitioned into smaller parts. The mind is indivisible. There might be different aspects of the mind, such as imagination, perception, or understanding. But Descartes argues that they are not properly speaking parts: when the mind perceives, it does so as a whole; there is not a fraction of the mind that is perceiving. The mind is a unified whole. Finally, Descartes can conclude that there is a real distinction between the mind and body. The mind is essentially an indivisible thinking thing, while the body is essentially extended and therefore divisible[33]:

> The first observation I make at this point is that there is a great difference between the mind and the body, inasmuch as the body is by its very nature always divisible, while the mind is utterly indivisible. For when I consider the mind, or myself in so far as I am merely a thinking thing, I am unable to distinguish any parts within myself; I understand myself to be something quite single and complete. Although the whole mind seems to be united to the whole body, I recognize that if a foot or arm or any other part of the body is cut off, nothing has thereby been taken away from the mind (Descartes, 1641/1996: 59).

Galileo provided a way to investigate the objective reality of the world. He told us that the world can be known for what it is and with mathematical rigor; the only price to pay is to confine qualities within the observer and outside of the observed physical bodies. Following in his footsteps, we started to understand the physical laws that rule the world, then came chemistry, biology, and social sciences. Through a combination of observations, measurements, logical inferences, and mathematical formalism, we did a good job of making sense of the natural world. Yet, we cannot disagree with Descartes' arguments: while everything we know about the physical world may be the

[33] Descartes concludes his Meditations by going back where he started: he concludes that he can now be confident about those things that he had cast into doubt in the First Meditation. The senses are normally adequate in guiding us around the world, and if we are in doubt, we can double-check our sensory perceptions with our intellect or our memory. He also notes that our memory can dispel the doubt presented in the Dream Argument. Any waking experience can be connected through memory to all other waking experiences, whereas in dreams, things happen in disconnected and somehow random manner.

product of a misled inference, the existence of our own mind is undoubtable. It is hard to believe that natural sciences are faulty and eventually will tell us everything we may want to know (and more) about the world. But at the same time, they do not seem to have space for our own subjective existence.

The tension between the two extremes generates this unique ontological problem whose implications are major for both our understanding of the world and our existence. Up until Galileo, the mind (i.e., the soul) was considered a central element of Aristotelian doctrine, a part of the world subject to the same rules. Contemporary philosophers of mind and cognitive scientists usually distance themselves from Descartes, as they frame him as the original sinner: he is guilty of having separated the mind from the body. However, after examining his arguments, one should be persuaded of their validity. It is not his arguments that are wrong, but one of the premises: the distinction between primary properties and secondary qualities.

9 – Conclusion

Since the beginning of the Early Modern era, we have witnessed the unstoppable rise of modern science. A quantitative science that made its success building on Galileo's premises: to study a phenomenon it is necessary that the observer abstracts from her own – *subjective* – point of view and embrace a sort of 'view from nowhere', capable of depicting reality in its true, *objective* form. Science's successes are a testament to the world – we can predict phenomena to such an accurate degree of precision that it would be unreasonable to negate the existence of an external and objective world. The coincidences would be just too many: science is too successful to negate its validity and the external world.

Conversely, through the cartesian *cogito ergo sum,* we must agree that the only certain and undoubtable piece of evidence we have is about our own individual and subjective existence. Moreover, upon reflection, we realize that everything else we know, we know it within consciousness. So, we should be at least open to the possibility that what we perceive about the world is subject to the 'rules' of our own minds. This realization is a further step from Galileo's distinction between primary properties and secondary qualities. Everything, including what Galileo thought to be primary properties, is subjectively experienced. Certainly, different subjects will experience them in a very similar way (and we may know it because, for example, they provide similar reports about it). But to take consciousness seriously means to take the subjective point of view of each individual as the primary source for any piece of

evidence about the world, including those that compose our best scientific theories. In other words, we realize that the view from nowhere is a construct of our mind, and it does not portray a picture of reality from a neutral point of view but merely reflects the conceptualization of that point of view from the perspective of individuals that share certain characteristics in terms of evolution, development, and education.

Before moving on to the next chapter, it can be useful to highlight some tenets that we have encountered so far.

-Phenomenal realism: not only is consciousness real, not an illusion, and exists but consciousness *equals* existence. We exist because we are conscious, and we know that we exist because we are conscious.

-Everything we know, we know within consciousness. Hence, consciousness is the upper bound of knowledge.

-Certain aspects of our subjective experience seem to be entirely subjective (e.g., the taste of fresh strawberries) while others seem to be entirely objective (e.g., the depth of a lake). However, both strawberries and the depth of the lake are subjectively experienced.

Finally, one question and a partial answer: in light of present considerations about consciousness, what should we do of science? If the scientific method is unfit to fit experience into our image of the world, should we discard it altogether? Of course not. Galileo opened the way for a very successful tool of inquiry. He showed that through observation and perturbation, predictions, and explanations we can understand an incredible range of phenomena. Moreover, we can describe models and test their consistency through what appears to be a universal language: mathematics.

Bibliography

Anstey, P. (2000). *The Philosophy of David Bolye*. Oxford: Routledge.

Barnes, J., ed. (1984). *The Complete Works of Aristotle*, Volumes I and II. Princeton: Princeton University Press, 1984.

Bianchi, M. (2014). Il dire galileiano per titoli: una nota lessicale su *Il Saggiatore*. *Zeitschrift für romanische Philologie*, 130, 802–814.

Campbell, K. (1984). *Body and Mind*. New York: Anchor Books.

Chalmers, D. (1996). *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D. (1995/2010). Facing Up to the Problem of Consciousness. In Chalmers, D., *The Character of Consciousness* (pp. 3–34). New York: Oxford University Press.

Churchland, P. (1985). Reduction, Qualia, and Direct Introspection of Brain States, *Journal of Philosophy*, 82, 8–28.

Cooper, J. M., ed. (1997). *Plato: Complete Works*. Indianapolis: Hackett.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G.P.

Dennett, D. (1978) Current Issues in the Philosophy of Mind. *American Philosophical Quarterly,* 15, 249–261.

Dennett, D. (1990) Quining Qualia. In *Mind and Cognition*, W. Lycan (ed.). Oxford: Blackwell (pp. 519–548).

De Santillana, G. (1955). *The Crime of Galileo*. Chicago University Press.

Descartes, R. (1641/1996). *Meditations on First Philosophy*. Cambridge University Press. Cambridge, UK.

Drake, S. (1957). *Discoveries and Opinions of Galileo*. Doubleday. Garden City, NY.

Foster, J. (1993). Dennett's Rejection of Dualism. *Inquiry*, 36, 1–2.

Frankish, K. (2017). *Illusionism: As a theory of consciousness*. Exeter, UK: Imprint Academic.

Galilei, G. (2017). *Il Saggiatore*. CreateSpace Independent Publishing Platform. Lexington, USA.

Goff, P. (2017). *Consciousness and Fundamental Reality*. New York, NY: Oxford University Press.

Goff, P. (2019). *Galileo's Error: Foundations for a New Science of Consciousness*. New York: Pantheon Books.

Graziano, M. S. A. (2019). *Rethinking Consciousness: A Scientific Theory of Subjective Experience*. W. W. Norton, New York.

Hatfield, G. (2018). René Descartes. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.). Retrieved from: <https://plato.stanford.edu/archives/sum2018/entries/descartes/>.

Hume, D. (1739/1922). *Treatise on Human Nature*. New York, Prometheus Books.

Lewis, C. S. (1990). Conscience and Conscious. In *Studies in Words*. Cambridge University Press

Lotze, H. (1884). *Lotze's System of Philosophy*. Bosanquet, B. (ed.). Clarendon Press: Oxford, UK.

James, W. (1879). The Spatial Quale. *J. Specul. Philos*, 1879: 13, 64–87.

Koch, C. (2020) *Re-enchanting the World: A Review of Galileo's Error: Foundations for a New Science of Consciousness by Philip Goff*. Book Review available online at https://cpb-us-e1.wpmucdn.com/sites.ucsc.edu/dist/0/158/files/2020/05/FinalKoch-review-of-Galileos-Error-20.pdf.

Koestler, A. (2014). *The Sleepwalkers: A History of Man's Changing Vision of the Universe*. London: Penguin Books.

Nagel, T. (1989). *The View from Nowhere*. New York: Oxford University Press.

Seth, A. (2016). The Real Problem. *Aeon Magazine*. Retrieved from: <https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one>.

Trape et al. (1965/2010) *Nuova Biblioteca Agostiniana. Opere di Sant'Agostino. Edizione latino-italiana*, 44 vols. Roma: Città Nuova Editrice, 1965–2010. Complete.

Uzgalis, W. (2020). John Locke. *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). Edward N. Zalta (ed.). Retrieved from: <https://plato.stanford.edu/archives/spr2020/entries/locke/>.

Van Gulick, R. (2018) Consciousness. *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Edward N. Zalta (ed.). Retrieved from: <https://plato.stanford.edu/archives/spr2018/entries/consciousness/>.

Woolhouse, R. S. (1971). *Locke's Philosophy of Science and Knowledge*. New York: Barnes and Noble.

Woolhouse, R. S. (1983). *Locke*. Minneapolis: University of Minnesota Press.

Woolhouse, R. S. (1988). *The Empiricists*. Oxford: Oxford University Press.

Chapter Three: The Axiomatic Approach of Integrated Information Theory

1 – Introduction

The mind–body problem is the problem of how to reconcile the ineffable mind with the tangible matter. Any true explanation of the mind presupposes an explanation for its qualitative character, the *what-is-likeness* of consciousness (Nagel, 1974). Whenever we try to address this problem in physical terms, as standard natural sciences prescribe, we face the hard problem of consciousness (Chalmer 1995), the problem that an explanation in physical terms will invariably leave out the subjective and qualitative character of consciousness. Physical explanations that start with the brain and its physical properties leave open an explanatory gap (Levine, 1983) between the physical properties of the brain and the phenomenal properties of consciousness. Is there a way out from this stalemate?

In the previous chapter, I argued that part of our misleading intuitions about the hard problem stems from the difficulty to reconcile the effectiveness that science had to provide objective knowledge about the physical world with the fact that the only thing that we know for sure is the existence of our own subjective experience. Here, I wish to show why the problem in these terms may be ill-posed. In this chapter, I will illustrate the axiomatic approach of integrated information theory (IIT) and how its adoption is a necessary step to dissipate the hard problem and solve the mind–body problem, thus putting consciousness and the physical back together in a common framework. I will first illustrate why we need a theory of consciousness. Then I will briefly present IIT and its axioms. Finally, I will address some criticisms that have been raised in recent years against IIT's axiomatic approach, showing why they are not relevant.

2 – Why we need a theory of consciousness

Investigating consciousness presents some critical challenges. Objective proxies for consciousness, such as behavioral, functional, and neural correlates can provide useful insights but do not yield definitive answers for those cases that diverge significally from the neurotypical human adult.

Consider how difficult it can be to determine the conscious state of an unresponsive brain-injured patient. Furthermore, what about newborns and babies that do not yet display a fully developed brain? What about other non-human mammals? The more we venture into the biological world, the less clear are the answers. Are reptiles conscious, or is their brain not developed enough? What about an octopus that does not even present a clearly defined brain and has neurons distributed all over its body, tentacles included? The same questions apply to artificial systems that have been developed and will be developed in the following years. In all these cases, the answer is the same: we cannot know without a principled way to address the question (Oizumi et al., 2014).

On the other hand, the most staggering feature of consciousness is its intrinsicality, i.e., the fact that my conscious experience exists for me. I am certain that my experience exists, and the very same moment I start doubting it, I am thus confirming its existence[34]. Nothing is more evident or directly accessible to me than my own experience. However, the very same property of intrinsicality makes it impossible for me to know with certainty that other beings are conscious. Strictly speaking, I do not need exotic cases like octopi; I cannot be sure that even my neighbor is not a zombie.

3 – The intrinsic perspective and the world as an inference

Many progresses have been made in science in terms of identities and structural explanations. For example, one may consider the discovery that water is an inorganic compound of one atom of hydrogen and two atoms of oxygen structured in a particular way. Similarly, identities between phenomena at different spatial and temporal scales have been used to explain phenomena that were apparently irreducible to one another (e.g., temperature and mean kinetic energy). Similar strategies have been used as attempts to explain consciousness.

Perhaps the most common example is the identity between the objectively measurable firing of C-fibers and the subjective feeling of pain. According to this thesis, pain *is* the firing of C-fibers. Needless to say, such a hypothesis, although suggestive, is

[34] See chapter 2.

barren: it does not matter how well the two events (the physical event of neuronal cell firing and the phenomenal event of feeling pain) correlate; the correlation itself does not explain *why* when the C-fibers' fire pain is present. Moreover, the properties of pain seem rather different from the properties of C-fiber, and in order to establish an identity, it would seem necessary that the properties of one are the properties of the other. However, this issue could be set aside as an instance of a temporary lack of knowledge about the brain and all its properties. After all, the connection between consciousness and the brain is evident, and to negate it would be negating common sense.

But why do we believe that explanations at the neuronal level should explain consciousness? More specifically, how do we justify this hypothesis? The general story tells us about an incredibly complex universe organized through hierarchies of scaffolded entities. This universe can be studied through the lenses of biology, chemistry, and physics, each domain at a different level of generality. But only physics really matters: at the bottom of reality, there are only fields and particles, or whatever the current best scientific theory describes as fundamental entities. A further hypothesis is added: only fundamental entities have genuine causal interactions, everything else at the higher levels of reality merely supervenes over these fundamental entities (Chalmers, 1996). Such a conception of the universe relies on the view from nowhere, the conceptualization of the universe that prescinds from our specific and subjective point of view (Nagel, 1986). There are two major problems with this view (i.e., the dominant view among philosophers and neuroscientists). The first problem is that it collides with my own experience of being a unitary being. But the view from nowhere is a fiction of the mind: it is a useful abstraction that we create to make sense of what our senses tell us when we perform observations and manipulations of reality. An *empirical* experiment presupposes a conscious experimenter by definition. Consciousness is hidden in plain sight all along. The difference between studying any natural phenomenon and consciousness is that in the former case, we can forget about our condition of the conscious observer and just assume it. In fact, phenomenon in Greek literally means "thing appearing to view," i.e., something that is experienced. But consciousness *is* experience. Consciousness is different from any generic natural phenomenon, because it is not a phenomenon. Its intrinsic nature presupposes that it cannot be experienced by others, and it only exists for itself. Therefore, contrary to any other scientific phenomenon when we study consciousness, we need to factor in its intrinsic nature.

The best way is to use phenomenology to retro-engineer consciousness and infer its physical requirements. Phenomenology is the study of conscious experience from within, which can be done through the tools of introspection and reason (Ellia et al., 2021). Starting with phenomenology is necessary to avoid the stalemate of the hard problem. In other words, following a rationalist tradition the idea is that for

consciousness being the way it is there must be a reason. This is the opposite of what is usually done: start from some plausible physical substrate, typically some interconnected set of neurons in the brain, and postulate that it would somehow give rise to experience (Tononi, 2015). Notably, this has already been argued convincingly in the 19th century. For example, Schopenhauer wrote:

> Materialism … tries to find the first and simplest state of matter, and then to develop all the others from it, ascending from mere mechanism to chemistry, to polarity, to the vegetable and the animal kingdoms. Supposing this were successful, the last link of the chain would be animal sensibility, that is to say cognition; which, in consequence, would then appear as a mere modification of matter, a state of matter produced by causality. Now if we had followed materialism thus far with clear notions, then, having reached its highest point, we should experience a sudden fit of the inextinguishable laughter of the Olympians. As though waking from a dream, we should all at once become aware that its final result, produced so laboriously, namely cognition was already presupposed as the indispensable condition at the very first starting-point, at mere matter. With this we imagined that we thought of matter, but in fact we had thought of nothing but the subject that represents matter, the eye that sees it, the hand that feels it, the understanding that knows it. Thus the tremendous p*etition principii* … Materialism is therefore the attempt to explain what is directly given to us from what is given indirectly. (Schopenhauer et al., 2018)

4 – From consciousness to its physical substrate

The goal is to examine the structure of experience through introspection and determine those properties that are present in every experience. These properties are determined according to some characteristics: they are evident, in the sense that they are directly available in my experience, are not filtered by other forms of reasoning, and do not need further proof except their presence. Such properties of experience are, strictly speaking, self-presenting and in no need of re-presenting. We express these properties by axioms. Moreover, the axioms should capture the minimum set of properties that apply to all experiences, in the sense that it is not possible to conceive an experience that does not present any of these properties. At the same time, it is always possible to conceive two different experiences that share only these five properties. Finally, any pairing of them should not involve contradiction, and no axiom should be derivable from the others. In other words, the axioms must be about experience, evident, essential, complete, consistent, and independent (Tononi, 2015). According to these criteria, the axioms of IIT are intrinsicality, composition,

information, integration, and exclusion (Tononi, 2015; Tononi et al., 2016; Tononi & Koch, 2015).

It must be noted how none of these properties dictates any specific content of experience. In this sense, they constitute the matrix of all possible experience without describing any experience in particular. They are the universal and invariant forms of consciousness; any specific content of experience must necessarily present additional properties to those described by the axioms (Ellia *et al.* 2021). Finally, each axiom is necessary, but only the five taken together are both necessary and sufficient (Tononi, 2015). Ideally, for any property that we may be tempted to include among the axioms, there will be a conceivable experience that does not present that property. For this reason, it is important to distinguish between essential properties (those defined by the axioms) and typical properties (those commonly present in experience but not essentially). For example, space feels extended, and time feels flowing; while time and space are typical of our experience of the world, they are not essential to it (Ellia *et al.* 2021; Haun and Tononi 2019).

The founding pillar of IIT is the equation of experience and existence[35]. Consciousness is real, it exists for me. Even if I hallucinate, my hallucinated experience is real, *for me*. Moreover, conscious experience is characterized by the presence of regularities. By regularities I mean, for example, the fact that when we close and reopen our eyes, the world is still there, or the fact that when I think about moving my arm, my arm moves, and it does not just move but moves in a certain way. In short, the world we experience seems to follow some rules. Every night, it ceases to exist, and every morning, it starts anew, as far as we are concerned. But we cannot avoid noticing how the world seems to go on even without us, even without us existing, at least from our own perspective. This ignites the suspicion that there is indeed a world that is not dependent upon us. These are the regularities that we want to explain when we are in search of a theory of consciousness (Tononi, 2015). However, we face two alternatives: solipsism or realism. Solipsism is the idea that everything that exists, is only my mind. Realism implies that there is a mind-independent world. But solipsism does not explain anything. In fact, solipsism is the assumption that there is not an explanation. Thus, we can accept its alternative, realism, on the ground that it is the best explanation to account for regularities within consciousness, given the two alternatives. Furthermore, within our experience, we attribute physical existence to what can affect and be affected by perturbations and manipulations[36]. Finally, we observe that what is real (mind-independent) and physical

[35] What follows in this section is due to Giulio Tononi, personal communication.

[36] This criterion of physical existence is also Eleatic Principle Grasso, M. (2019). IIT vs. Russellian Monism: A Metaphysical Showdown on the Content of Experience. *Journal of Consciousness Studies*, *26*(1-2), 48-75. , Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. https://doi.org/10.4249/scholarpedia.4164

(can be observed and manipulated) can also be partitioned and studied in its components (methodological reductionism).

Phenomenal axioms can be operationalized in cause-effect terms. This means that there is a way to talk about mind-independent entities by partitioning and perturbing their components and observing how these manipulations affects the whole and its parts. In other words, we conceive physical existence as cause–effect power. Hence, we determine physical postulates as an inference to the best explanation for the axioms: if experience is intrinsic, structured, specific, unitary and definite so must be its physical substrate. In other words, we can postulate the condition of possibility for each axiom in operational terms. If experience equals existence, the five axioms together are the condition of possibility of existence, and the postulates are the condition of possibility for the axioms in operational terms, then the postulates account for existence in cause–effect terms. In short, by affecting the operational translation, we get a non-phenomenological language to express the existence we initially discovered within phenomenology – after all, this is what translations are for. We can now look into the world, described in cause–effect terms, and search for entities that match our description – entities that exist intrinsically and are composite, informative, integrated, and exclusive (as defined by the postulates). If our guided search turns out successful, we will have found an intrinsic entity, i.e., one that exists for itself, one that is the subject of its existence. And since experience equals existence, "being an intrinsic entity, properly defined, is one and the same thing as being conscious" (Tononi 2017). As per postulates, an intrinsic entity specifies a maximally irreducible cause–effect structure (MICS). There is a relevant choice of terminology here: axioms are the indubitable starting point of our reasoning (as much as it is absolutely certain that, by experiencing, we are), while postulates are in fact postulated. Finally, the identity is an explanatory identity in the sense that translating (always within consciousness) phenomenal properties into operational language allows us to explain regularities within experience and make inferences about other potential conscious entities. But being explanatory does not weaken the ontological status of our identity, nor does it reduce it to a mere isomorphism. The explanatory power comes from the fact that the identity is an ontological one: there are not two things, there is only one (the thing that exists), which can be defined in both phenomenal and operational terms, but, as our knowledge of the world is bound by our own experience, a phenomenal description will always be the most fundamental one in both our epistemology and ontology.

Summing up, the indubitable truth that experience is existence has led us to find by reflection those universally necessary and sufficient properties for something to be an experience, namely intrinsicality, composition, information, integration, and exclusion. On one hand, they make possible all experience by capturing its quintessential form; on the other hand, if they are sufficient for something to be

experience, they are not sufficient for something to be the experience that it is, i.e., specifically, since nothing exists in general. Then, by assuming an experience-independent reality, whose physical existence means having decomposable cause–effect power, we can translate the properties of phenomenology in operational terms, i.e., in cause–effect language. We thereby get a sufficient reason for experience being as described by the axioms: its cause–effect power is intrinsic, compositional, specific, integrated, and maximally exclusive. If this is a sufficient reason, the ground of this reason is not within the postulates. What makes possible this double description of existence, meaning in both phenomenological and causal terms, is the fundamental identity between experience and maximally irreducible, specific, structured, and intrinsic cause–effect power. Once this is in place, the possibility of knowing is secured. The identity serves as a blueprint to make inferences about what we know less (the world) based upon what we know best (our own experience). Moreover, it allows us to carve the world at its joints, establishing what is an intrinsic entity and what is not.

5 – A primer on integrated information theory

IIT employs a unique epistemology in the landscape of modern neuroscientific approaches to consciousness: a *phenomenology-first approach*. Therefore, rather than starting from neural mechanisms and inevitably facing the hard problem, IIT begins with phenomenology and then infers the mechanisms of consciousness (Tononi et al., 2016).

As such, the theory puts forward five *axioms* derived from reflection on our consciousness, meant to capture the essential properties of every conceivable experience. They describe the structure of consciousness, its very fabric, so no experience can fail to satisfy these properties (Ellia et al., 2021). In IIT, these essential properties are *intrinsicality*, *composition*, *information*, *integration*, and *exclusion*. By the axioms, a conscious experience is: (i) intrinsic: it exists for its own subject and not for an external observer; (ii) structured: it is composed by phenomenal distinctions bound by relations; (iii) specific: it is informative by being the particular way it is; (iv) unitary: it is an integrated whole, not reducible to any of its parts (distinctions and relations); (v) definite: it has borders and is definite in content; it contains what it contains, neither less nor more (Haun & Tononi, 2019; Oizumi et al., 2014; Tononi, 2015; Tononi et al., 2016).

To each axiom corresponds a *postulate*, which in conjunction describe the ontological (causal) properties of the physical substrate of consciousness. Briefly, postulates are an operationalization in cause–effect terms of the axioms and provide the causal reasons for which experience is as the axioms describe it. Postulates are expressed in mathematical language. The postulates are intrinsicality, composition, information, integration, and exclusion. IIT aims to explain consciousness in terms of the *integrated information* of a physical substrate (Oizumi et al., 2014; Tononi, 2015). A substrate is defined as a system of connected units in a state (e.g., a set of neurons firing or not firing in a brain). Simply put, integrated information quantifies the causal power of the system and its parts upon themselves. For a given physical system in a state, its integrated information is assessed by *unfolding* its cause–effect structure, which in turn captures how the system in that state constrains its past and future states. To unfold the cause–effect structure of a system means to partition and perturb its element in all possible ways following a compositional approach (Albantakis & Tononi, 2019). The cause–effect structure obtained through this process can be represented as a abstract simplicial complex (Haun & Tononi, 2019; Maaten & Hinton, 2008).

According to the postulates, for a physical substrate to underlie experience, it must have *intrinsic*, *compositional*, *specific*, *integrated*, and *maximal cause-effect power*. Therefore, a proper conscious substrate specifies a maximally irreducible conceptual structure (MICS) or more generally, a cause-effect structure[37]. IIT goes on to posit a fundamental *identity* between an experience and the cause–effect structure of the physical substrate. The structure of a particular conscious experience is identical to a MICS:

> The maximally irreducible conceptual structure (MICS) generated by a complex of elements is identical to its experience. The constellation of concepts of the MICS completely specifies the quality of the experience (its quale '*sensu lato*' (in the broad sense of the term). Its irreducibility Φ^{Max} specifies its quantity. The maximally irreducible cause-effect repertoire (MICE) of each concept within a MICS specifies what the concept is about (what it contributes to the quality of the experience, i.e., its quale '*sensu stricto*' (in the narrow sense of the term), while its value of irreducibility φ^{Max} specifies how much the concept is present in the experience. (Oizumi *et al.* 2014)

The cause–effect structure is described in information-theoretic terms, with integrated information Φ quantifying its irreducibility. Notably, candidate physical substrates of consciousness are characterized from a topological rather than functional point of

[37] IIT's terminology evolved through time. Such a structure can also be called Q-structure or conceptual structure or, in general, cause–effect structure. See Appendix B.

view, and this allows for two functionally identical systems to be phenomenologically distinct. This means that two systems can be given the same set of inputs to provide the same set of outputs, and yet they can present radically different phenomenological properties (Grasso et al., 2021; Oizumi et al., 2014). Moreover, as long as their causal structures are identical, different systems can have the same experience (multiple realizability).

The essential properties of experience described by the axioms and operationalized through the postulates can account for the presence or absence of consciousness in a given system. If a system has a cause-effect structure, then the system is conscious (Oizumi et al., 2014; Tononi, 2015). Moreover, since each property of the cause–effect structure reflects a property of the phenomenal structure, specific experiences can be characterized in cause–effect terms (Ellia et al., 2021; Haun & Tononi, 2019).

Consciousness is *being*, not *doing*; therefore, consciousness does not have a function and cannot have an immediately evident adaptive value. However, integrated systems, such as those that IIT deems conscious, are more efficient in terms of available functions per number of elements. Therefore, given the constraints of space and energy that determine the evolutionary trajectory of every organism, a complex organism should sustain complex conscious experiences. In fact, computational models within the framework of IIT suggest that, for an organism, the exposure to environments richer in complexity may lead to an increase in internal connectivity and richer intrinsic cause–effect structures (Albantakis, 2020a; Albantakis et al., 2014; Albantakis & Tononi, 2015; Juel et al., 2019).

IIT makes a vast and diversified set of predictions that can in principle falsify the theory if disconfirmed by empirical evidence (Tsuchya *et al.* 2020). Intuitively, Φ should be high when consciousness is present and low or zero when it is minimal or absent. Two main predictions follow from this: a) brain areas that constitute the physical substrate of consciousness should have high Φ, while brain areas that do not contribute to consciousness should have minimal Φ; b) brain areas that constitute the physical substrate of consciousness should have high Φ when consciousness is present and minimal Φ when consciousness is absent (Ellia et al., 2021; Tononi et al., 2016). These predictions have been investigated both experimentally and with computational models of the brain (Balduzzi & Tononi, 2008; Tononi et al., 2016). Given the computational intractability of Φ for real world systems, proxy measures and heuristics are necessary. A crucial prediction of IIT, the breakdown of effective cortical connectivity during dreamless sleep, was confirmed through a combination of transcranial magnetic stimulation and high-density EEG (Massimini et al., 2005). Later, a quantitative measure, the perturbational complexity index (PCI), was developed (Casali et al., 2013). PCI quantifies integration and segregation, two properties that are necessary for high value of integrated information within a physical substrate (Tononi et al., 2016). Finally,

studies in healthy subjects during wakefulness, dreamless and dreaming sleep, and general anesthesia indicate that the loss and recovery of consciousness are associated with the breakdown and recovery of the capacity for information integration in the corticothalamic system (Casarotto et al., 2016; Tononi et al., 2016). Based on its theoretical apparatus, IIT predicts that the full NCC must be organized anatomically in a way that is ideally suited to support high values of integrated information. Therefore, proponents of the theory predict that the cortical area, known as the *posterior hot zone*, constitutes the physical substrate of consciousness[38] (Koch et al., 2016; Tononi et al., 2016).

-

6 – The 'Core' of Integrated Information Theory

6.1 Intrinsicality

Axiom of intrinsicality: " […] every experience is subjective—it is for the subject of experience, from its own intrinsic perspective, rather than for something extrinsic to it" (Haun & Tononi, 2019).

An entity exits intrinsically if and only if it exists *for itself,* i.e., its existence does not depend upon an external observer experiencing it. IIT rephrases the Cartesian *cogito*: *I experience, therefore I am* (Oizumi et al., 2014). Consciousness being intrinsic means that experience is real and that it exists for the subject who experiences it. I am immediately and absolutely certain of the existence of my experience, and, as an entity, it does not need the existence of external observers: it exists *for me*, its subject. The subjective character of consciousness is widely recognized as one of the central aspects of consciousness (Nagel, 1974).

Postulate of intrinsicality: " […] intrinsicality means that a candidate substrate of consciousness, such as a set of neurons in the cerebral cortex, must have cause-effect power upon itself, rather than just with respect to sensory inputs and motor outputs" (Haun & Tononi, 2019).

[38] Note that, while neuroscientists generally speak of neural correlates of consciousness, IIT makes a stronger constitutive claim, justified by its fundamental identity. The set of neurons that IIT deems necessary for consciousness do not merely correlate their activity with its presence: Their cause–effect structure is identical to one's subjective experience. IIT also predicts that the physical substrate of consciousness is not necessarily immutable, but it can "move" around the brain according to how effective connectivity and other factors affect its causal structure Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450-461. https://doi.org/10.1038/nrn.2016.44 .

The criterion of physical existence in IIT is given by the Eleatic Principle: to exist means to have cause–effect power (Grasso, 2019; Tononi, 2015). To exist intrinsically requires a further constraint, to have cause–effect power upon itself. To exist from its own intrinsic perspective, independent of extrinsic factors, requires that the system's mechanisms in their current states "make a difference" to the probability of some past and future state of the system (Tononi & Koch, 2015).

6.2 Composition

Axiom of composition: "[…] every experience is structured, being composed of phenomenal distinctions and relations" (Haun et al., 2017).

The *axiom of composition* states that consciousness is structured, in the sense that experience is composed of various phenomenal distinctions bound by relations (Tononi 2015). One may consider the experience of staring at New York's skyline. Within my visual field, I can distinguish a left and a right side, top and bottom, but also shapes, colors, and shadows. I can count many skyscrapers and, for each building, countless windows. Each building has a size, shape, and color. The sky is blue and extended, and so on. I can isolate countless phenomenal distinctions within my visual field. Moreover, distinctions are not just on their own, but they are related to other distinctions (i.e., structured) in a specific way. While IIT's way to characterize the structure of experience is quite specific, the idea that experience is structured has a long tradition in phenomenology (see for example (Kant, 1998)).

Postulate of composition: "Composition means that one must consider the structure of intrinsic cause-effect power—how various combinations of neurons can have causes and effects within the system (causal distinctions) and how these distinctions overlap causally (causal relations)" (Haun & Tononi, 2019).

As much as phenomenal experience is structure, so must be the physical substrate of consciousness. Moreover, to investigate a physical system as potential physical substrate of consciousness, a causal compositional approach is necessary (Albantakis & Tononi, 2019). In fact, neither a reductionist approach (considering only first order elements) or a holistic one (considering the system as a whole) will suffice. Instead, it is necessary to consider each potential subset of elements within the system to determinate its causal effect within the system. In other words, subsets of system elements (composed in various combinations) must have cause–effect power upon the system (Tononi, 2015; Tononi & Koch, 2015).

6.3 Information

Axiom of information: "Every experience is specific: it is the particular way it is" (Haun & Tononi, 2019).

The axiom of information states that consciousness is informative or specific, i.e., any experience has the particular character it has. While this axiom may seem almost trivial, it is one of the most overlooked aspects of consciousness. Indeed, while other axioms have been somehow recognized through the history of philosophy, information is one of the novelties of IIT. Information highlights the incredible richness and complexity of our experience (Haun et al., 2017). We may consider for example one of the simplest conceivable experiences: staring at a white, empty canvas. Ignoring any other thought, feeling, or sensation, even such a simple experience has a massively complex structure (Ellia et al., 2021; Haun & Tononi, 2019; Haun & Tononi, In preparation). The mere experience of an extended space requires millions and millions of phenomenal relations to *inform*[39] its specific phenomenal character. Thus, a visual experience of complete darkness and Michelangelo's *The Last Judgement* are both very informative in this sense.

Postulate of information: "Information means that the causes and effects specified by various combinations of neurons are specific states of specific subsets of neurons, yielding a specific cause-effect structure" (Haun & Tononi, 2019).

Since every experience is specific, is the way it is, and not in any generalized way, so must be its physical substrate. The cause–effect structure of the conscious system must be a specific set of specific cause–effect repertoires in a specific state, thereby differing in its specific way from other possible structures (Tononi & Koch, 2015). For a mechanism in a state, its cause–effect repertoire specifies the probability of all possible causes and effects. A cause–effect repertoire characterizes in full the cause–effect power of a mechanism within a system by making explicit all its cause–effect properties. To determine a cause–effect repertoire, one must perturb the system in all possible ways, in order to determine how a mechanism in its present state makes a difference to the probability of the past and future states of the system (Oizumi et al., 2014; Tononi, 2015). The cause–effect structure is the set of cause–effect repertoires specified by all subsets of system elements related in a certain way. Finally, the notion of information in IIT differs greatly from Shannon's information (Oizumi et al., 2014).

[39] The Latin etymological sense of "information" is "to give form," i.e., to order and structure. Moreover, this reflects the fact that information in physical terms in IIT is doubly associated with causation: information is causal, and causation is informative Albantakis, L., Ellia, F., & Tononi, G. (In preparation). .

6.4 Integration

Axiom of integration: "Integration means that every experience is unified, being irreducible to independent components." (Haun & Tononi, 2019).

The axiom of integration suggests that consciousness is unitary, i.e., every experience cannot be reduced to its components. In other words, experience is a non-decomposable whole (Oizumi *et al.* 2014, Tononi *et al.* 2016). Your visual experience right now is not just the left side of your visual field plus the right side; both sides are indeed present, they compose your experience, but the experience itself is more than the sum of its components. Kant and Descartes famously argued for the unity of consciousness (Descartes, 1984; Kant, 1998).

Postulate of integration: "Integration means that causal distinctions and relations, as well as the overall cause-effect structure they compose, only exist if they are irreducible if they cannot be reduced to independent causes and effects." (Haun & Tononi, 2019).

Since every experience is unitary whole, so must be the cause-effect structure. A unitary cause-effect structure is beyond and above its parts, being irreducible to its components. Therefore, the cause-effect structure exist as a single entity, despite being composed by individual components.

6.5 Exclusion

Axiom of exclusion: "Exclusion means that every experience is definite—it contains what it contains, neither less nor more."

The axiom of exclusion states that experience has borders, and it is definite in content (Tononi 2015; Tononi *et al.* 2016). Content-wise, my experience right now contains only the phenomenal distinctions and relations present in it, neither less (a subset) nor more (a superset). At any given time, experience has the set of phenomenal distinctions it has and nothing more or less. It is in this sense that consciousness is exclusive: at any given time, there is only one experience rather than a superposition of multiple partial experiences.

Postulate of exclusion: "Exclusion means that causal distinctions and relations, as well as the cause-effect structure they compose, must be definite, containing what they contain—neither less nor more. What defines the set of neurons that constitute the physical substrate of consciousness as opposed to any of its subsets or supersets—is being maximally irreducible, as measured by integrated information." (Haun & Tononi, 2019).

6.6 Identity

> Integrated information theory proposes an explanatory identity between a particular experience and the particular cause-effect structure specified by a physical substrate in its current state (Haun & Tononi, 2019).

Given the phenomenology first approach, the identity is given by stipulation. After determining the essential characteristics of the phenomenal structure (the axioms), the characteristics of the cause-effect structure are determined to explain those phenomenal properties. Hence, this identity is an explanatory identity (Haun and Tononi 2019). More importantly, not differently from axioms and postulates, the identity is posited a priori. Therefore, IIT's fundamental identity is radically different from the identity between water and H2O, which necessity is established a posteriori (Chis-Ciure & Ellia, 2021).

The a priori identity has important consequences for the theory and the mind-body problem. Besides avoiding the conceivability scenarios, it also prevents the hard problem. In fact, starting from phenomenology, we twisted the problem upside-down, and we do not need to explain the phenomenal in terms of the physical, but vice-versa: through the explanatory identity we are able to explain the physical in terms of the phenomenal. Moreover, the predictions made by the theory (e.g., that the neural correlates of consciousness are a global maxima of integrated information) can be independently tested in an empirical way. If confirmed, IIT then solves the mind-body problem not only theoretically but also empirically, effectively providing a principled answer to the question: "How can we fit our subjective experience into our objective description of the world?"

6 – Conclusion

In this chapter I explored how the axiomatic approach of IIT prevents the hard problem and solves the mind-body problem. Through the characterization of the essential properties of experience, and their translation in physical terms, IIT twist the mind-body problem upside down, successfully providing a solution which is empirically testable. This explanatory project is carried out by the fundamental identity between the structure of phenomenal experience and the cause-effect structure of the physical substrate of consciousness in a state. Moreover, the identity is an identity a

priori. Therefore, it is not limited by the explanatory gap and it prevents conceivability scenarios.

Bibliography

Albantakis, L. (2020a). Integrated information theory. In J. M. Morten Overgaard, Asger Kirkeby-Hinrup (Ed.), *Beyond Neural Correlates of Consciousness*. Routledge.

Albantakis, L. (2020b). Unfolding the Substitution Argument. *Conscious(ness) Realist*. https://www.consciousnessrealist.com/unfolding-argument-commentary/

Albantakis, L. (2021). My conscious(ness) biases. *Conscious(ness) Realist*. https://www.consciousnessrealist.com/consciousness-biases/

Albantakis, L., Ellia, F., & Tononi, G. (In preparation).

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing

Complexity. *PLOS Computational Biology*, *10*(12), e1003966. https://doi.org/10.1371/journal.pcbi.1003966
Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, *21*(5), 459. https://www.mdpi.com/1099-4300/21/5/459
Albantakis, L., & Tononi, G. (2015). The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. *Entropy*, *17*(8), 5472-5502. https://www.mdpi.com/1099-4300/17/8/5472
Albantakis, L., & Tononi, G. (2019). Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy*, *21*(10), 989. https://www.mdpi.com/1099-4300/21/10/989
Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.
Balduzzi, D., & Tononi, G. (2008). Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLOS Computational Biology*, *4*(6). https://doi.org/10.1371/journal.pcbi.1000091
Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy*, *23*(3), 362. https://www.mdpi.com/1099-4300/23/3/362
Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), 18803. https://doi.org/10.1038/s41598-020-75943-4
Barnes, J. (1984). *The Complete Works of Aristotle* (Vol. I and II). Princeton University Press.
Berkeley, G. (1734). *The analyst: A discourse addressed to an infidel mathematician*. Wilkins, David R.
Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, *5*(198). https://doi.org/10.1126/scitranslmed.3006294
Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., G Casali, A., Trimarchi, P. D., & Boly, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of neurology*, *80*(5), 718-729.
Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200-219.
Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9-10), 6-61.
Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press.
Chis-Ciure, R., & Ellia, F. (2021). Facing up to the Hard Problem of Consciousness as an Integrated Information Theorist. *Foundations of Science*. https://doi.org/10.1007/s10699-020-09724-7
Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. *Journal of Philosophy*, *78*(2). https://doi.org/10.5840/jphil198178268
Churchland, P. S. (1986). *Neurophilosophy: Toward A Unified Science of the Mind-Brain* (Vol. 97). MIT Press.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8). https://doi.org/10.1016/j.tics.2011.06.008

Cooper, J. M. (1997). *Plato: Complete Works*. Hackett.

Crick, F. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul* (Vol. 37). Scribners.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in Neuroscience*, *2*, 263-275.

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529-14534. https://doi.org/10.1073/pnas.95.24.14529

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Descartes, R. (1984). *Meditations on First Philosophy*. Caravan Books.

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *12*(2), 41-62. https://doi.org/10.1080/17588928.2020.1772214

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*. https://doi.org/10.1016/j.concog.2019.04.002

Ellia, F. (2020). Francis Crick and the Hard Problem of Consciousness. *Sistemi Intelligenti*, *2*(August), 267 - 286. https://doi.org/10.1422/96324

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, J., Haun, A., Albantakis, L., Boly, M., & Tononi, G. (2021). Consciousness is a structure, not a function. *In preparation*.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2007). Masking Disrupts Reentrant Processing in Human Visual Cortex. *Journal of Cognitive Neuroscience*, *19*(9), 1488-1497. https://doi.org/10.1162/jocn.2007.19.9.1488

Ferrell, B. A. (1950). Experience. *Mind*, *LIX*(234), 170-198.

Feyerabend, P. (1975). *Against Method* (Vol. 87). New Left Books.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11-12), 11-39.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127-138. https://doi.org/10.1038/nrn2787

Friston, K. (2013). Consciousness and Hierarchical Inference. *Neuropsychoanalysis*, *15*(1), 38-42. https://doi.org/10.1080/15294145.2013.10773716

Gomez, J. D., Mayner, W. G. P., Beheler-Amass, M., Tononi, G., & Albantakis, L. (2021). Computing Integrated Information (Φ) in Discrete Dynamical Systems with Multi-Valued Elements. *Entropy*, *23*(1), 6. https://www.mdpi.com/1099-4300/23/1/6

Grasso, M. (2019). IIT vs. Russellian Monism: A Metaphysical Showdown on the Content of Experience. *Journal of Consciousness Studies*, *26*(1-2), 48-75.

Grasso, M., Haun, A., & Tononi, G. (2021). Of maps and grids. *Submitted to this issue*.

Hanson, J. R., & Walker, S. I. (2019). Integrated Information Theory and Isomorphic Feed-Forward Philosophical Zombies. *Entropy*, *21*(11), 1073. https://www.mdpi.com/1099-4300/21/11/1073

Hansson, S. O. (2006). Falsificationism Falsified. *Foundations of Science*, *11*(3), 275-286. https://doi.org/10.1007/s10699-004-5922-1

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy*, *21*(12). https://doi.org/10.3390/e21121160

Haun, A., & Tononi, G. (In preparation). *Do you see all the dots?*

Haun, A. M., Tononi, G., Koch, C., & Tsuchiya, N. (2017). Are we underestimating the richness of visual experience? *Neuroscience of Consciousness*, *3*(1).

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257. https://doi.org/https://doi.org/10.1016/0893-6080(91)90009-T

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8

Huxley, T. H. (2011). *Collected Essays*. Cambridge University Press.

Juel, B. E., Comolatti, R., Tononi, G., & Albantakis, L. (2019). When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. ALIFE 2019: The 2019 Conference on Artificial Life,

Kant, I. (1998). *Critique of Pure Reason*. Cambridge University Press.

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab001

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5). https://doi.org/10.1038/nrn.2016.22

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions* (Vol. 2). University of Chicago Press.

Kuhn, T. S. (1970). Logic of Discovery or Psychology of Research? In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 22). Cambridge University Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (Vol. 4, pp. 91-195). Cambridge University Press.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494-501. https://doi.org/10.1016/j.tics.2006.09.001

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365-373. https://doi.org/10.1016/j.tics.2011.05.009

Laudan, L. (1983). The Demise of the Demarcation Problem. In R. S. Cohen & L. Laudan (Eds.), *Physics, Philosophy and Psychoanalysis: Essays in Honor of Adolf Grünbaum* (pp. 111--127). D. Reidel.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Leibniz, G. W. (1714). *Monadology*.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(October), 354-361.

Linassi, F., Zanatta, P., Tellaroli, P., Ori, C., & Carron, M. (2018). Isolated forearm technique: a meta-analysis of connected consciousness during different general anaesthesia regimens. *British Journal of Anaesthesia*, *121*(1). https://doi.org/10.1016/j.bja.2018.02.019

Maaten, L. V. D., & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.
Mahner, M. (2007). Demarcating Science from Non-Science.
Mallatt, J. (2021). A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness. *Entropy*, *23*(6), 650. https://www.mdpi.com/1099-4300/23/6/650
Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., & Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, *309*(5744). https://doi.org/10.1126/science.1117256
Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: A toolbox for integrated information theory. *PLOS Computational Biology*, *14*(7), e1006343. https://doi.org/10.1371/journal.pcbi.1006343
Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science*, *372*(6545), 911-912. https://doi.org/10.1126/science.abj3259
Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*(October), 435-450.
Nagel, T. (1986). *The View From Nowhere* (Vol. 37). Oxford University Press.
Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*, *19*(5), 979-996.
Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLOS Computational Biology*, *10*(5). https://doi.org/10.1371/journal.pcbi.1003588
Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science*, *313*(5792). https://doi.org/10.1126/science.1130197
Popper, K. (1935). *The Logic of Scientific Discovery*. Routledge.
Putnam, H. (1991). The `corroboration' of theories. *Philosophy of Science*, 121--137.
Revonsuo, A. (2009). *Consciousness: The Science of Subjectivity*. Psychology Press.
Ruse, M. (1977). Karl Popper's philosophy of biology. *Philosophy of Science*, *44*(4), 638-661.
Sanders, R. D., Tononi, G., Laureys, S., & Sleigh, J. W. (2012). Unresponsiveness ≠ Unconsciousness. *Anesthesiology*, *116*(4). https://doi.org/10.1097/ALN.0b013e318249d0a7
Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, Adenauer G., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., & Massimini, M. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, *25*(23). https://doi.org/10.1016/j.cub.2015.10.014
Schopenhauer, A., Welchman, A., Norman, J., & Janaway, C. (2018). *Schopenhauer: The World as Will and Representation: Volume 2*. Cambridge University Press.
Seth, A. (2016). The real problem. *Aeon*.
Singer, P. N. (2016). Galen. *The Stanford Encyclopedia of Philosophy*, *Winter 2016 Edition*. <https://plato.stanford.edu/archives/win2016/entries/galen/>
Sprigge, T. (1971). Final causes. . *Proceedings of the Aristotelian Society*, *45*.
Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. https://doi.org/10.4249/scholarpedia.4164

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450-461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140167. https://doi.org/doi:10.1098/rstb.2014.0167

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*, 102877. https://doi.org/https://doi.org/10.1016/j.concog.2020.102877

Vanhaudenhuyse, A., Charland-Verville, V., Thibaut, A., Chatelle, C., Tshibanda, J.-F. L., Maudoux, A., Faymonville, M.-E., Laureys, S., & Gosseries, O. (2018). Conscious While Being Considered in an Unresponsive Wakefulness Syndrome for 20 Years %U https://www.frontiersin.org/article/10.3389/fneur.2018.00671. *Frontiers in Neurology*, *9*(671 %M), %7 %8 2018-August-2028 %2019 Original Research %# %! Conscious while being considered in an unresponsive wakefulness syndrome for 2020 years. %* %<. https://doi.org/10.3389/fneur.2018.00671 %W %L

Veselis, R. A. (2006). The remarkable memory effects of propofol. *British Journal of Anaesthesia*, *96*(3), 289-291. https://doi.org/10.1093/bja/ael016

Watson, J. B., & McDougall, W. (1929). *The battle of behaviorism: An exposition and an exposure*. W.W. Norton & Co.

Chapter Four: The Unfolding argument

1 – Introduction

The unfolding argument is an argument that aims to disprove the methodology of the so-called causal structure theories of consciousness in neuroscience. More specifically, it intends to show that IIT is either false or unscientific. Originally proposed in 2019 by Doerig and his colleagues, the argument has generated several responses from opposite sides (Albantakis, 2020b; Doerig et al., 2021; Hanson & Walker, 2019; Kleiner & Hoel, 2021; Negro, 2020). In general, causal structure theories assert that a target system is conscious if and only if its parts interact in a certain way, i.e., if it displays the causal structure deemed necessary by the theory's proponents (Doerig et al., 2019). Among causal structure theories, Doerig and colleagues include recurrent processing theory (RPT) and integrated information theory (IIT). A key element of criticism that the proponents of the unfolding argument move against IIT is that causal structure theories such as IIT postulate a dissociation between cognitive functions and phenomenal experience (Doerig et al., 2021; Doerig et al., 2019). This argument is particularly relevant from a philosophy-of-science standpoint, as it shed light on two important and often ignored aspects of theory building: the role of intuitions and the (pre-theoretical) assumptions on which a theory is built. Moreover, the unfolding argument has implications for the empirical testability of a given theory of consciousness.

For example, RPT's proponents claim that recurrent processing is both necessary and sufficient for consciousness (Lamme, 2006); visual experience occurs after the non-conscious first feedforward information processing re-enters the pre-activated neural circuits in a recurrent top-down way. It has been empirically observed that after subjects reported a lack of visual experience of masked stimuli, no recurrent activity was spotted in V1 (Fahrenfort et al., 2007).

IIT prominently focuses on the causal structure of the target system (Albantakis, 2020a; Oizumi et al., 2014; Tononi, 2015; Tononi et al., 2016). In particular, IIT

proposes an identity between the cause-effect structure of a system of elements in a state (for example, a set of neurons in a human brain) and its experience. Importantly, the identity is between the experience and the cause-effect structure, not the physical substrate per se (Tononi, 2015). A cause-effect structure is composed by maximally irreducible[40] distinctions bound by maximally irreducible relations. The irreducibility of a distinction can be quantified by φ, which is a measurement that captures how the distinction in its state at the present time constraints the other elements of the system in past and future states (Barbosa et al., 2021). The irreducibility of a relation can be quantified by $\varphi_{relation}$, which is a measurement that captures the way in which the same set of elements overlaps (Haun & Tononi, 2019). Distinctions can be thought of as the building blocks of a cause-effect structure, and relation as the way they are structured. Finally, the irreducibility of a cause-effect structure itself can be quantified by Φ (Oizumi et al., 2014; Tononi et al., 2016). IIT's proponents hypothesize that if a system has a maximally irreducible cause-effect structure quantified by Φ^{max}, then the system is conscious[41] and the Φ^{max}-value captures the "quantity" of its consciousness. Finally, according to IIT's proponents, the neural correlates of consciousness are the areas in brain that present the highest value of integrated information (Tononi et al., 2016). A candidate area, due to the presence of grid-like neurons which connectivity should maximize integrated information, is the posterior hot zone (Koch et al., 2016). Importantly, IIT does not characterize consciousness as what the brain *does* but as how the cause-effect structure of the physical substrate of consciousness *is* (Ellia et al., 2021; Grasso et al., 2021).

2 – The unfolding argument

Nominally, the unfolding argument challenges both RCP and IIT, as well as other theories that propose a dissociation between functions and consciousness, but IIT is its primary target as the title of the paper suggests (Doerig et al., 2019). Proponents of the unfolding argument claim that causal structure theories should be disregarded on the grounds that, in general, a relevant aspect of causal structure theories is that they require a particular kind of architecture, rich in feedback connections, rather than a specific function. However, for any physical system that has feedback connections, it is possible to find a different physical system (arguably with a different corresponding causal structure) that is functionally indistinguishable. Moreover, proponents of the unfolding argument point out that any experiment that investigates the physical substate of consciousness ultimately is expressed as some kind of

[40] See chapter 3.
[41] Notably, due to the formalism of the theory, the Φ value of a system cannot be negative.

function, while the neural architecture, being the object of inquiry, cannot be used to guide the research. Hence, Doerig and colleagues argue that IIT is either incoherent or unscientific. In their view, IIT might be incoherent because it should accept that non-feedback systems can be conscious since they would fare exactly as the recurrent ones in an experimental setting. It may be unscientific because if IIT's proponents bite the bullet and claim that only recurrent systems can be conscious, then the claim cannot be falsified and should therefore be disregarded as belonging outside the scope of science. In what follows, I will present the argument with more details, and in the following sections, I will point out some flaws of its flaw and its inapplicability within the context of the neuroscientific research.

Both recurrent neural networks and feedforward neural networks are Krohn-Rhodes function approximators (Hornik, 1991; Hornik et al., 1989), i.e., an input-output function can be approximated to any degree of accuracy (Doerig et al. 2019). Therefore, for a recurrent network performing a certain input-output function, there is an equivalent feedforward network that performs the exact same input-output function (LeCun et al., 2015).

Therefore, any behavioral experiment can be interpreted as an input-output function, and the same function can, in principle, be realized within a recurrent or feedforward network. Moreover, the same function can be realized within an indefinite number of different networks, as the function per se does not depend upon the structural properties of the network. For example, a masked stimulus is shown to the subject, who she has to press a button if she sees it, or not press the button if she does not. This experiment can be rendered as an input (the stimulus being shown or not) and an output (the subject pressing the button or not), and it can be expressed with a countless number of both recurrent and feedforward networks (Doerig et al. 2019).

In a more schematic way, the unfolding argument goes as follows:

> (p1): Science relies on physical measurements.
>
> (p2): For any recurrent system with an input-output function, there exists a feedforward system with the same input-output function, and vice versa.
>
> (p3): Two systems that have identical input-output functions cannot be distinguished via any experiment that relies on physical measurements (other than a measurement of brain activity itself or of the other internal workings of the system).
>
> (p4): We cannot use measures of brain activity as a-priori indicators of consciousness, because the brain basis of consciousness is what we are trying to understand in the first place.
>
> (c): Therefore, either causal structure theories are falsified (if they accept that unfolded networks can be conscious) OR they are

> outside the realm of scientific inquiry (if they maintain that unfolded feedforward networks are not conscious despite being empirically indistinguishable from functionally equivalent recurrent networks) (Doerig et al. 2019).

Looking at each step of the argument in more detail, Doerig and colleagues assert that the scientific study of consciousness should rely on physical measurements (p1), but such measurements cannot be based on neural correlates per se, because that is what we want to explain in the first place (p4). We can render any experiment in terms of input-output functions, but the same function can be implemented by both feedforward and feedback architectures (p2), making the two kinds of systems indistinguishable from the standpoint of an input-output experiment (p3). Therefore, they conclude that either IIT is falsified or unfalsifiable and therefore unscientific.

To summarize, Doerig and colleagues claim that all theories that rely on causal structures as a general means to explain consciousness are threatened in that they are either false or unfalsifiable, regardless of how they define the desired causal structure. Due to its prominence, IIT is the main target of the unfolding argument. Doerig and colleagues also noticed how the unfolding argument does not affect other approaches that do not rely on causal structure, namely the global neural workspace (Baars, 1993; Dehaene et al., 1998), higher order theory (Lau & Rosenthal, 2011), and predictive processing (Friston, 2010, 2013). The main reason for this is that these theoretical approaches propose models of consciousness in which what is relevant is the *function* performed by a system, not the way in which such a function is implemented within the system. In other words, the unfolding argument does not apply to functionalist theories of consciousness and applies only to causal structure theories[42] (Doerig et al., 2021; Doerig et al., 2019; Tsuchiya et al., 2020).

3 – What is wrong with the unfolding argument

The unfolding argument raises many concerns. First, its applicability to the biological world is questionable. As noted by Mallatt, "Perhaps the unfolding argument could apply in some idealized world that is based only on logic, but not in the dangerous and competitive world of reality" (Mallatt, 2021). Feedforward networks are less

[42] This claim has recently been challenged by Kleiner and Hoel, who illustrated how the unfolding argument can be extended to a more general problem of falsification that applies to any non-trivial model of consciousness Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab001 .

efficient when compared with integrated ones. For biological systems that evolve under spatial and energetic constraints, it is unlikely to develop purely feedforward architectures (Koch, 2019). The higher efficiency of integrated networks within complex environments was also observed in simulation studies (Albantakis, 2020a; Albantakis et al., 2014). However, Hanson and Walker proved that small networks (thus, not real-world biological systems) can also implement the same function over the same amount of node (Hanson & Walker, 2019). Albantakis (2020b) also noticed how the unfolding argument does not need the substitution of an integrated network with a feedforward one, but only the possibility of such a substitution.

Besides the issue of pertinence to real-world systems, the unfolding argument still presents three major criticalities. I will refer at them as 1) originality; 2) the radical concept of consciousness; and 3) the shallow concept of science.

3.1 – Original contribution to the debate

A first question that is necessary to ask is – what novelty does the unfolding argument bring to the debate in consciousness studies? Arguably, very little. The unfolding argument can be considered in its two major components: the functional equivalence between integrated and feedforward systems and the fact that theories that postulate a dissociation between functions and consciousness cannot be tested.

The first horn of the problem shows that there is a functional equivalence between integrated systems (as formally defined by IIT) and feedforward systems. Such equivalence can be expressed in terms of the function performed; given the same inputs, the two systems will provide the same output. Notably, this functional equivalence was already illustrated by the proponents of IIT five years prior to the publication of the unfolding argument (Albantakis et al., 2014; Oizumi et al., 2014). Further, within the context of IIT, this equivalence is justified by theoretical claims (Ellia et al., 2021; Oizumi et al., 2014). Moreover, Grasso and colleagues show how two functionally equivalent networks, an integrated "grid" and a feedforward "map", are not equivalent when it concerns the explanation of the subjective character of space, an aspect that further justify the double dissociation between functions and consciousness (Grasso et al., 2021; Haun & Tononi, 2019).

The second horn of the problem demonstrates that if a theory postulates a dissociation between functions and consciousness, then such a theory cannot be tested scientifically. This idea is however not original, as such a radical version of functionalism had already been proposed by Cohen and Dennett (Cohen & Dennett,

2011). In their famous paper, Cohen and Dennett conceive a thought experiment called "the perfect experiment," which aimed to disprove the class of theories that postulate separate correlates of consciousness for consciousness and cognition (cognitive functions). More specifically, Cohen and Dennett intended to show that any neurobiological theory based on the division between experience and function cannot be empirically confirmed or falsified and is thus outside the scope of science (Cohen & Dennett, 2011). Doerig and colleagues' goal was the same: to argue against the empirical tractability of theories that postulate the dissociation between consciousness and functions. This is a radical version of functionalism that, following Negro, I will call "input-output functionalism" (Negro, 2020).

Under the light of originality, the unfolding argument does not fare well. It relies on the functional equivalence between feedforward networks and integrated ones, something that was put forward many years ago by the proponents of IIT. Moreover, this functional equivalence is used to argue against theories that dissociate consciousness from cognitive functions, in other words, the unfolding argument is a "perfect experiment" under disguise.

3.2 – A radical concept of consciousness: input-output functionalism

To understand the implications of the unfolding argument, perhaps we need to take a step back and consider what a scientific theory of consciousness ought to explain. In the previous chapters, I defined consciousness as experience. Therefore, a subject is conscious if there is something that feels like being that subject (Nagel, 1974). This position is often called "phenomenal realism." The mere presence of qualities is sufficient for consciousness, not the ability to attend or report said qualities or even the display of intelligent behavior. The explanandum[43] of a scientific theory of consciousness should be, unsurprisingly, consciousness itself and not the behavior we commonly associate with it (Ellia, 2020). Therefore, to explain consciousness, we need to account for its presence and qualitative character (Ellia et al., 2021).

IIT's proponents highlight how the misconception about the explanandum of a scientific theory of consciousness leads to the so-called fallacy of misplaced objectivity. The fallacy is in assuming that science ought to explain objective things in an objective way. Committing this fallacy leads to the conclusion that consciousness

43 What a theory should explain, or its explanatory target

can be studied scientifically only on its behavioral, functional, and neural correlates, leaving experience outside of the picture (Ellia et al., 2021).

Consider a generic experiment: when the experimenter asks the subject if she sees the stimulus, the experimenter does not (or at least should not) care about the verbal reports of the subject except her subjective experience. In other words, it should be clear that seeing the stimulus is different from reporting the stimulus (Ellia et al., 2021). Unfortunately, we cannot access other people's experience, so we need to rely on different kinds of proxies to make inferences about their state of consciousness. Once again, it is important to emphasize that consciousness itself is what we ultimately care about:

> When we consider a "subjective report about consciousness", are we taking consciousness itself as evidence, or not? We maintain that the "ground truth" data in consciousness science are conscious experiences (Tsuchiya et al. 2020).

The intrinsic[44] nature of consciousness makes it a unique case in the taxonomy of scientific *explananda*. One may be tempted to say that consciousness is by definition non-observable. Additionally, contrary to any other natural phenomena, my experience cannot be subject to multiple and independent observations[45]. Yet, subjectively, I am aware of my own consciousness. This kind of first-person phenomenological constraints are crucial.

After having clarified what the explanandum for a theory of consciousness is, we can revisit the unfolding argument to consider the assumptions on which it stands. In particular, consider (p1): "Science relies on physical measurements." In a broad sense, this is undeniable. However, the assumptions on which this claim rests are ambiguous. Tsuchiya and colleagues (2020) highlight how (p1), when considered in the context of cognitive neuroscience, is open to two interpretations: one which claims that physical measurements are necessary to do science (and seems trivial) and the other which claims that physical measurements are all that we should be concerned about when we do science. Doerig and his colleagues seem to indicate that to investigate consciousness, only behavioral evidence is relevant[46]. Such assumptions imply a certain meta-scientific view: methodological behaviorism (Tsuchiya et al. 2020).

---

44 According to IIT, one of the five essential properties of consciousness is intrinsicality Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. https://doi.org/10.4249/scholarpedia.4164 . See chapter 3 for more details.

45 In chapter 1 I discussed this point from the opposite angle: consciousness differs from any other entity because we cannot observe it from a neutral point of view. Moreover, every observation, including scientific observations, are 'bounded' by our own consciousness.

46 As shown in section 1, Doerig and colleagues believe that to observe the neural substrate during through experiments that should verify a theory of consciousness equals to raising the question in a similar fashion to Cohen and Dennett Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8).

> Broadly, this stance effectively takes private mental events as incompatible with empirical science. Thus, a psychological science of any sort – including consciousness science – would rely strictly on the observation of behaviors by a subject (and their relation to stimuli), without imputing any further non-observable properties to the subject. Of course, a consciousness researcher must be imputing some non-observable properties to their subjects – Doerig et al. hope to study consciousness, after all – but a behaviorist consciousness researcher would consider such qualities to be empirically irrelevant, and perhaps even theoretically invalid (Tsuchiya et al. 2020).

Negro refines this claim further and highlights how the position endorsed by Doerig and colleagues should be called input-output functionalism. Moreover, Negro shows through a textbook example, that the unfolding argument is limited in scope (Negro, 2020).

The functionalist label is more appropriate than the behavioristic one. Historically, behaviorism represented the attempt to naturalize psychology by getting rid of any concept inherent to the mind, treating it like a black box whose inputs and outputs could be studied through behavior (Watson & McDougall, 1929). In contrast, functionalism was the attempt to open the black box of behavior to unveil, through the study of the internal cognitive architecture, i.e., the study of functions that connect an input *y* with an output *x*.

An input-output functionalist could be thought of as someone who believes that functional equivalence corresponds to equivalence in terms of consciousness, because functional states and consciousness states are not dissociable within this context. If pressed, an input-output functionalist should admit that if a system *x* is functionally equivalent to a human being, then *x*'s state of consciousness is identical to that of the human being in question. It is easy to understand this position by imagining that if an input-output functionalist builds a Sophia 2.0, a robot that can behave indistinguishably from another human being, then the input-output functionalist will consider Sophia 2.0 conscious. Notably, causal structure theories of consciousness will be silent on the state of consciousness of Sophia 2.0 until they are able to assess its causal structure[47].

We live in a world rich in complexity, and most of our behavior is dictated by assumptions that we make and hold true about our environment. Descartes proved that the only thing that I can know with absolute certainty is the fact that I exist and

---

https://doi.org/10.1016/j.tics.2011.06.008 , Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*. https://doi.org/10.1016/j.concog.2019.04.002 .

[47] Or by appealing to a more general rule, e.g., according to IIT, von Neumann architecture – used in contemporary computers – is almost certainly unfit to sustain conscious experience.

that to exist means to be conscious[48]. Yet, this certainty does not imply that I am compelled into assuming solipsism. I usually act (and believe) as though other people are conscious too. Moreover, a life of observations of other people's behavior and my personal experience led me to establish certain correlations that I generally find reliable in my everyday life. For example, when a person talks or interacts with her environment, it means that the person in question is conscious. If a person is sleeping, I am tempted to say that she is not conscious; however, thanks to my own experience, I know that sometimes it is possible to be asleep and conscious at the same time when dreaming. Therefore, the disconnection from the environment and the inability to interact with it do not always imply the absence of consciousness. Likewise, even though my phone can perform marvelous tasks and even answer my questions through its vocal assistant, I do not believe that it is conscious. These are just some of the many examples that demonstrate how I can make reliable inferences that allow me to understand and interact with the world in my everyday life. However, inferences cannot rely purely on intuitions, but must be guided by consistent principles. For this reason, we need a theory and cannot rely only on behavioristic or functional intuitions.

Clinical evidence suggests that functional evidence may not be the best choice when we investigate consciousness. Since antiquity, anesthetics have allowed for medical procedures of all sorts. As the word suggests, an anesthetic drug prevents the patient from feeling her environment. We can say that the main purpose of anesthetics is to prevent the patient from experiencing the surgery. This goal can be achieved in different ways with different drugs. For example, a proper dosage of ketamine keeps patients conscious, but they hallucinates and feel disconnected from their environments, effectively unable to experience the medical procedure (Sarasso et al., 2015). However, most drugs work in a different way; they just render patients completely unconscious. One of the most common among these anesthetics is propofol. This drug is used every day in thousands of hospitals across the world, and it causes the loss of consciousness in the patient, who falls into a deep, dreamless sleep, not only unable to experience her environment, but also – as far we know – unable to experience anything at all, like in dreamless sleep. Unfortunately, this is not always what happens. Approximately one patient in every 1000 reports after the surgery that they were conscious all the while and could see and hear the surgeon performing the surgery, meanwhile unable to move any muscle of their body. To avoid any spontaneous movement, patients are also given drugs that paralyze their muscles for the duration of the procedure. Albeit imprisoned in their own bodies, we now have evidence that some patients stay awake and are not disconnected from their environment during surgery (Linassi et al., 2018; Sanders et al., 2012). Luckily, anesthetics are also known to have a common side effect: amnesia. While this limits

[48] See Chapter 2.

the damages, it also helps keep the phenomenon obscure; it is estimated that a much higher percentage than the reported 0.1% of patients may be present during surgeries. To prevent this, the anesthetist can use the "isolated forearm technique" which helps prevent complete paralysis by isolating the forearm and allowing the surgeon and patient to communicate (if necessary) through the contraction of the hand. The isolated forearm technique is considered the golden standard in consciousness monitoring during surgeries (Veselis, 2006).

Consider how difficult it is to distinguish between patients that are normally suffering from the so-called locked-in syndrome, which causes the total paralysis of the body, from patients that have the unresponsive wakefulness syndrome (UWS). Both cases lack entirely of behavior, but the UWS presents a disorder of consciousness while the locked-in patient is aware. Cases have been described in the literature where patients with the locked-in syndrome have been misdiagnosed for as long as 20 years (Vanhaudenhuyse et al., 2018).

A fairly recent approach, the so-called "tennis paradigm" (Owen et al., 2006) prove how the lack of normal behavioral evidence is not an evidence for the lack of consciousness. The approach goes as follows. During the experiment, a brain-injured patient, believed to be affected by the UWS, was put inside a fMRI scan and asked some questions. The patient was instructed to think about playing tennis if the answer was "yes" and to think about moving in her house if the answer was "no," and the brain activity was monitored. The subject and experimenter were able to have a meaningful conversation, leading to the hypothesis that the subject was conscious. It was only later that the expression "minimally conscious" was used to describe patients in similar conditions. Unfortunately, this protocol is not easy to apply, for example, a patient may no longer be able to hear, or understand language. Moreover, if the patient has been in the same condition for a long time, he or she may be severely depressed and hence lack the motivation to answer.

Despite the recent advances in our understanding of consciousness disorders, accurate diagnosis of severely brain-damaged patients is still a major clinical challenge. Standardized behavioral evidence without the support of neuroimage is not enough to make a precise diagnosis (Vanhaudenhuyse et al. 2018).

Since the aforementioned example of Sophia 2.0 is rather abstract and serves more as an intuition pump rather than providing us with a solid criterion for concrete real-world cases, consider how the input-output functionalist would address the following situation. Imagine two patients who have spent the last 20 years in a hospital bed, both showing minimal interaction with their environment, but none considered responsive. Solely based on behavior or their reaction to stimuli (input-output), the input-output functionalist is forced to say that either, both, or none of them is

conscious. In fact, ex hypothesis, for the input-output functionalist, if two systems (in this case, the two patients) perform the same function (in this case, none), then the two systems must share the same state of consciousness. In other words, unable to detect the differences between these two cases, the input-output functionalist can only conclude that whatever their state of consciousness is, is the same for both of them. However, let's imagine that a few weeks later, one of the two patients recovers. He is finally able to communicate, and talk share his terrible experience: he was conscious all the time, while completely unable to move. He was misdiagnosed with the UWS, while it should have been considered a case of the locked-in syndrome. Unfortunately, misdiagnosis of disorders of consciousness are frequent [add references], and one of the benefits of having a verified theory of consciousness is that it should help to make better diagnosis. How can the input-output functionalist reply to the recovering patient without looking at the cause-effect structure of the brain? She is either forced to say that the now-recovered patient is hallucinating and *believes* that he was conscious when he was in his previous condition, when in fact he was not. However, this seems very arbitrary and inconsistent with multiple patient reports. Alternatively, she has to assume that the patient was indeed misdiagnosed, but even if that's the case, the other patient is currently conscious and unable to act. Moreover, the input-output functionalist will not use brain activity as an indicator of consciousness if not exclusively from a functional point of view. In the case of Sophia 2.0, the causal structure-theorists, on the other hand, would not have an answer ready. His or her inference on the state of consciousness of the two patients would be guided by the information available about the causal structure of the brain.

Now, consider a similar but slightly different scenario. There are three patients, all of them apparently unresponsive and disconnected from their environment. Once again, behavioral correlates of consciousness are absent. This time, however. we can apply the tennis paradigm, and we observe the following: patient number one is able to have a conversation with the experimenters, and we then observe her brain activity. Patient number two is non-responsive to the tennis paradigm. His brain is active, but from a neurophysiological standpoint, there is not much overlap between patient one's and patient two's brain activity. Finally, patient number three is also non-responsive; he is not able to communicate through the tennis paradigm, but his brain activity shows similarities with the brain activity registered in patient one. How can the input-output functionalist interpret these results? Should she rely more on the behavioral aspects (e.g., being able to communicate with the experimenters) or try to establish a correlation in terms of brain activity instead?

From whatever angle one tries to frame it, the input-output functionalist lacks principled approaches to controversial cases within the clinical context, which is what we ultimately aim for. Beyond the philosophical and scientific interests in providing an explanation of consciousness, we need a theory, because we want a set of guidelines

that will help us make inferences about controversial cases that diverge from the neurotypical adult human brain.

3.3 – The problem with science

Finally, there is an important aspect of the unfolding argument regarding the particular way in which its proponents demarcate a scientific practice from an unscientific one.

The conclusion of the unfolding argument is that causal structure theories are either false or unfalsifiable, where the latter option is taken to imply that an unfalsifiable theory is non-scientific. Therefore, according to the unfolding argument's proponents, if the unfolding argument is sound, IIT, being a causal structure theory, is either false or non-scientific. Doerig and colleagues posits that a theory's proposition about a phenomenon is scientific only if it makes clear which conditions would falsify the proposition. IIT makes numerous and sometimes counter-intuitive predictions. Some of IIT's predictions are probably untestable, even in principle. For example, IIT claims that a simple eight-nodes system with certain well-defined properties can have a conscious experience and that experience is a spatial one (Haun & Tononi, 2019). In fact, Doerig and colleagues do not simply "demarcate" science from pseudo-science on the grounds of falsification; they expect that all statements derivable from a scientific theory should be testable. This seems unreasonably strict and it is in clear contraposition with how science works. Tsuchiya and colleagues noticed how untestable predictions are controversial although frequently encountered in science, for example, they cite Everett's many-worlds interpretation of quantum mechanics; relativistic accounts of "what it's like to fall into a black hole"; and the existence of gravitons in some versions of quantum gravity (Tsuchiya et al., 2020).

How to *demarcate* scientific practice from non-scientific practice is one of the oldest problems in the philosophy of science. Negro (2020) notes how Doerig's interpretation of the demarcation problem is not distant from a classic form of Popperian falsificationism (Popper, 1935). Notably, Doerig and colleagues do not provide any arguments for it. However, while common among scientists, falsificationsim has been largely criticized within the philosophy of science, and ultimately rejected. There are many, notable names in the ever-growing list of those who criticized falsificationsim, including (Feyerabend, 1975; Hansson, 2006; Kuhn, 1962; Kuhn, 1970; Lakatos, 1970; Laudan, 1983; Mahner, 2007; Putnam, 1991; Ruse, 1977).

Under a less rigid criterion, it is easy to see how IIT does not fall outside the scope of science. Most of IIT's predictions are testable in an empirical way. Consider one of the earliest empirical confirmation of IIT (Massimini et al. 2005). In the study, Massimini and colleagues observed the breakdown of cortical effective connectivity during sleep. The study was conducted by observing through high density EEG how cortical dynamics were affected by a transcranial magnetic stimulation. During wakefulness, after the initial local response at the site of the perturbation, a series of waves propagated in other areas. During NREM sleep, the initial local response was stronger but did not propagate and faded away shortly after (Massimini et al. 2005). It is easy to notice how the opposite results would have falsified the hypothesis that physical integration (through effective cortical connectivity) is necessary for consciousness in humans.

Moreover, IIT makes many other predictions that are testable in principle. For example, IIT can be falsified if the value of integrated information present in a brain is zero[49]. If the value is low and the subject is conscious, or the value does not change while the subject changes state of consciousness, then the theory is falsified. Similarly, IIT predicts that generalized epileptic seizures (generally associated with the loss of consciousness) will present lower value of Φ despite the high level of activity due to synchronization (Tononi, 2015). IIT also predicts that the brains of split-brain patients will include not one, but two separated areas that constitute a maxima of integrated information (Tononi, 2015). More generally, IIT predicts that the neural correlates of consciousness[50] are constituted by the area that corresponds toa maxima of integrated information. Specifically, it predicts that the global maxima of integrated information in a neurotypical human brain is located in the posterior hot zone (Koch et al., 2016; Tononi et al., 2016). Independent studies can falsify or verify this prediction. Indeed, such a project is currently being carried out by six independent research groups (Melloni et al., 2021). how such a precise prediction can be disproved by any independent study. Finally, psychophysical experiments should be able, to falsify the claims of the theory regarding the qualitative aspects of consciousness. If the cause-effect structure changes due to stimulations, but subjects do not report any difference in their experience, then the theory is falsified (Haun & Tononi, 2019).

49 Notice however that this option is unfeasible at the present time due to computational limitations. However, proxy measures have been proposed and approximations of Φ have been calculated for the human brain.

50 More specifically, while 'neural correlates of consciousness' is the common expression, IIT calls that 'the physical substrate of consciousness'. As the ontology of the theory implies an identity between the causal structure of the physical system and the experience.

4 – Conclusion

In this chapter I illustrated how the unfolding argument presents some criticalities in terms of originality, commitment to a radical definition of consciousness and to an obsolete concept of science. Notably, there is a difference between criticizing an argument and its underlying assumptions. The fact that radical functionalism and falsificationism are problematic views does not immediately disqualify the argument itself.

However, is important to bring attention to one's own premises. IIT makes its premises explicit by committing to an axiomatic approach. Explicit premises expose the theory to criticism, especially when those premises are not shared by the vast majority of the scientific community (as, for example, the need to start with phenomenology). But clear premises are necessary for internal consistency. Ultimately the value of axioms, as much as in consciousness science as in geometry, is to have a clear and well-defined starting point for our inquiry. Certain starting points are better than other, or at least better fit for a specific task. In this sense, both radical functionalism and falsificationism seem bad premises for study consciousness and demarcate scientific research from non-scientific research.

Bibliography

Albantakis, L. (2020a). Integrated information theory. In J. M. Morten Overgaard, Asger Kirkeby-Hinrup (Ed.), *Beyond Neural Correlates of Consciousness*. Routledge.

Albantakis, L. (2020b). Unfolding the Substitution Argument. *Conscious(ness) Realist*. https://www.consciousnessrealist.com/unfolding-argument-commentary/

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLOS Computational Biology*, *10*(12), e1003966. https://doi.org/10.1371/journal.pcbi.1003966

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy*, *23*(3), 362. https://www.mdpi.com/1099-4300/23/3/362

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8). https://doi.org/10.1016/j.tics.2011.06.008

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529-14534. https://doi.org/10.1073/pnas.95.24.14529

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *12*(2), 41-62. https://doi.org/10.1080/17588928.2020.1772214

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*. https://doi.org/10.1016/j.concog.2019.04.002

Ellia, F. (2020). Francis Crick and the Hard Problem of Consciousness. *Sistemi Intelligenti*, *2*(August), 267 - 286. https://doi.org/10.1422/96324

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, J., Haun, A., Albantakis, L., Boly, M., & Tononi, G. (2021). Consciousness is a structure, not a function. *In preparation*.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2007). Masking Disrupts Reentrant Processing in Human Visual Cortex. *Journal of Cognitive Neuroscience*, *19*(9), 1488-1497. https://doi.org/10.1162/jocn.2007.19.9.1488

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127-138. https://doi.org/10.1038/nrn2787

Friston, K. (2013). Consciousness and Hierarchical Inference. *Neuropsychoanalysis*, *15*(1), 38-42. https://doi.org/10.1080/15294145.2013.10773716

Grasso, M., Haun, A., & Tononi, G. (2021). Of Maps and Grids. *Submitted to this issue*.

Hanson, J. R., & Walker, S. I. (2019). Integrated Information Theory and Isomorphic Feed-Forward Philosophical Zombies. *Entropy*, *21*(11), 1073. https://www.mdpi.com/1099-4300/21/11/1073

Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, *21*(12). https://doi.org/10.3390/e21121160

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257. https://doi.org/https://doi.org/10.1016/0893-6080(91)90009-T

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab001

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5). https://doi.org/10.1038/nrn.2016.22

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494-501. https://doi.org/10.1016/j.tics.2006.09.001

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365-373. https://doi.org/10.1016/j.tics.2011.05.009

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Linassi, F., Zanatta, P., Tellaroli, P., Ori, C., & Carron, M. (2018). Isolated forearm technique: a meta-analysis of connected consciousness during different general anaesthesia regimens. *British Journal of Anaesthesia*, *121*(1). https://doi.org/10.1016/j.bja.2018.02.019

Mallatt, J. (2021). A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness. *Entropy*, *23*(6), 650. https://www.mdpi.com/1099-4300/23/6/650

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*(October), 435-450.

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*, *19*(5), 979-996.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10*(5). https://doi.org/10.1371/journal.pcbi.1003588

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting Awareness in the Vegetative State. *Science*, *313*(5792). https://doi.org/10.1126/science.1130197

Sanders, R. D., Tononi, G., Laureys, S., & Sleigh, J. W. (2012). Unresponsiveness ≠ Unconsciousness. *Anesthesiology*, *116*(4). https://doi.org/10.1097/ALN.0b013e318249d0a7

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, Adenauer G., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., & Massimini, M. (2015). Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine. *Current Biology*, *25*(23). https://doi.org/10.1016/j.cub.2015.10.014

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. https://doi.org/10.4249/scholarpedia.4164

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450-461. https://doi.org/10.1038/nrn.2016.44

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*, 102877. https://doi.org/https://doi.org/10.1016/j.concog.2020.102877

Vanhaudenhuyse, A., Charland-Verville, V., Thibaut, A., Chatelle, C., Tshibanda, J.-F. L., Maudoux, A., Faymonville, M.-E., Laureys, S., & Gosseries, O. (2018). Conscious While Being Considered in an Unresponsive Wakefulness Syndrome for 20 Years %U https://www.frontiersin.org/article/10.3389/fneur.2018.00671. *Frontiers in Neurology*, *9*(671 %M), %7 %8 2018-August-2028 %2019 Original Research %# %! Conscious while being considered in an unresponsive wakefulness syndrome for 2020 years. %* %<. https://doi.org/10.3389/fneur.2018.00671 %W %L

Veselis, R. A. (2006). The remarkable memory effects of propofol. *British Journal of Anaesthesia*, *96*(3), 289-291. https://doi.org/10.1093/bja/ael016

Watson, J. B., & McDougall, W. (1929). *The battle of behaviorism: An exposition and an exposure*. W.W. Norton & Co.

Appendix A – Is IIT an Emergentist Theory?

Introduction

As a theoretical notion, emergence has a long and tortuous history, and as such there is no generally accepted theory of emergence. It is nowadays often discussed in many fields of philosophical and scientific inquiry such as biology and neuroscience, and more generically in the complex system literature. The etymological root of the word comes from the Latin "*emergere*", i.e., to rise. In very broad terms, *emergence* is essentially characterized by a certain *connection* with its constituents or precursors, and by *novelty* with respect to them, features that can be viewed from an epistemological and/or ontological point of view. The former is usually linked with the problem of the unpredictability of emergent phenomena based on knowledge of their parts; the latter generally hints at a real feature that emerges independently of our knowledge of its constituents. The diversity of accounts of emergence is generated and justified by the variety of phenomena deemed as emergent: from quantum entangled states, covalent bonding, traffic jams, phase transitions, stigmergy to consciousness and economic relations.

A related doctrine is that of *generative atomism*, according to which everything is generated from atoms and combinations of them. Facts about atoms coupled with facts about their rules of composition entail every fact about composite entities – they have no novelty with respect to their atomic constituents. 'Atom' here does not denote the entity described by physics as an 'atom', but whatever entity is taken to be fundamental, i.e. non-composite, in our universe (e.g. electrons, neutrinos, quarks etc.). For an entity to be an atom its must be (i) immutable (thus indivisible), i.e. unchangeable, and (ii) individually distinguishable (Humphreys 2016). *Generative atomism physicalism* is the metaphysical view according to which fundamental entities are physical atoms and everything else is determined by these atoms and their configurations. The methodological component of generative atomism entails that any system can be (ideally) reconstructed from the parts and their rules of structured arrangements: one level or domain is *reduced* to another. Via such a reduction we get an ontology of only fundamental entities (i.e. microphysicalism), everything else being derivative. This view is at odds with the idea that there are properties of the whole which cannot be reduced to its elements, generally considering features such agency and consciousness as something closer to an illusion rather than a real feature of the world.

It is not surprising that emergence raised to prominence in the field of complex systems science and its preceding scientific milieu, where a breakdown of the generative atomism *desideratum* happened in multiple, interrelated moments. For instance, the inherent uncertainty of certain properties of subatomic components in quantum mechanics posed a major challenge to classical determinism and its conception of predictability. The recognition of non-linearity gave a further blow to the notion of predictability even in deterministic settings. Then, with the advent of complex systems, not only components, but their interaction become subjected to uncertainty, by having characteristics like time-dependence, contextuality, non-linearity (Thurner et al. 2017).

In his landmark *More Is Different* (1972), Nobel laureate physicist Philip Anderson emphasizes how emergent phenomena, products of increased complexity at different scales, undermine the dream to reconstruct all higher levels from simple fundamental laws. Even though all matter obeys simple electrodynamics and quantum theory, Anderson illustrates the "constructionist fallacy" of generative atomism with an example of broken symmetries in many-body physics as generators of emergent behavior, distributed hierarchically as we increase complexity at each stage of organization. The main idea is that complex macroscopic objects cannot be understood *only* in terms of simple extrapolation of the properties of some set of particles; instead, there is a *layering* of complexity, where at each level entirely new properties appear, and a new kind of level-specific fundamentality arises. "At each stage entirely new laws, concepts, and generalizations are necessary, requiring

inspiration and creativity to just as great a degree as in the previous one […]." (Anderson 1972).

Thus, the conceptual relation between emergence and generative atomism is that the justification for the first is usually a failure of the latter. Therefore, a property is *emergent* with respect to some system if and only if none of its components possess that property, i.e. it is *novel*. Not only properties can emerge; objects, states, events, processes or laws can be products of emergence. Most of the literature converges on the following core features for any example of emergence: (i) relationality or micro-macro effect, i.e. the fact that any item emerges *in relation* to previous items or condition and not in isolation; (ii) novelty, i.e. the emergent item is or possesses some property not possessed by its components or precursors (Bunge, 2003; De Wolf & Holvoet, 2005; Humphreys, 2016). Other more peripheral features of emergent phenomena that often occur are autonomy, holism or systemic character or whole-coherence (Kauffman, 1993), dynamical evolution (i.e. emergents arise diachronically) (Hooker, 2011), robustness and flexibility, decentralized control, top-down or downward causation (Ellis 2011). One omission here is the property of unexplainability, which gives rise to a form of strong, 'spooky' or mystical emergence, rejected by most theorists. Moreover, it is worth mentioning that emergent properties are by definition non-aggregative, in the sense that they are "more (or less) than the sum of their parts", which is a reason why emergence typically does not occur in an interesting way in aggregates, but in complex systems. The focus on interaction in the latter gives a substantial ground for appealing to emergence, whereas aggregates in the sense of the summing of properties of the components at the level of the whole, e.g. weight, extension, volume etc., are not considered genuine cases of emergent behavior.

Recently, Humphreys (2016) provided a two-dimensional categorization of emergence. On the first dimension, there are *ontological*, *inferential*, and *conceptual* forms of emergence, each with its characteristics. On the second dimension, there are *synchronic* and *diachronic* emergence. Combined, these yield six distinct types of emergence. The *inferential* approach to emergence takes the novel phenomena to be underivable from the properties of and relations between their constituents. *Conceptual* emergence demands that the conceptual apparatus that describes or captures an emergent feature is not part of the theoretical framework (i.e. concepts, law statements, theories) used to represent the entities that give rise to it. Finally, *ontological* emergence entails that what emergently arises is a genuine, mind-independent feature of the world. In their *diachronic* form, all these types of emergence hold that the emergent phenomena arise over time as the system evolves, while *synchronic* emergence generally describes phenomena of pattern formation.

Given the above, a natural question arises: is IIT an emergentist theory? Are there according to the theory entities or properties that emerge from others? In fact, there are at least three cases where this question could be legitimately asked:

- Do Macro Units emerge from Micro Units?
- Do Higher-order Mechanisms emerge from First-order Mechanisms?
- Does the Cause-Effect Structure emerge from the Main Complex?

In the following we will examine each of these cases.

Macro Units & Micro Units

IIT predicts that the physical substrate of consciousness (PSC) is a global maximum of integrated information (Tononi et al. 2016). We call this "main complex". The brain is constituted by interacting networks of brain sub-systems, which are constituted by local networks of neurons, which are constituted by columns of neurons, which are constituted by neurons, which are constituted by atoms and so on. We can observe and manipulate all of them in order to find the elements that constitute the main complex. The main complex can be represented as a set of interconnected elements in a state, where the state of the system depends on the states of its individual elements. We call these elements "intrinsic constituents", as they are the units that determine the intrinsic mechanisms. However, the intrinsic constituents themselves can be represented as a set of interconnected elements in a state, where the state of the system (the intrinsic constituent) depends on the states of its individual elements. In this case we call Macro Units the former and Micro Units the latter, in such a way that a Macro element is a set of Micro elements (Marshall et al. 2018). Recall that IIT adheres to a criterion of parsimony enforced by a causal principle of exclusion (which prevents causal overdetermination, i.e., multiple causes for a single effect). Now the question is, do Macro Units emerge from Micro Units?

Once again it is important to consider how IIT deals with the ontological criterion: consciousness is existence, and consciousness (hence existence) is defined by the axioms in phenomenological terms and by the postulates in operational terms. Therefore, if we have a *maximum of integrated information* at the level of Macro Units, Micro Units are excluded. They do not exist for themselves, and this is reflected by Exclusion, implying that they have a lower level of integrated information when compared to Macro Units. In fact, changes in the states of Micro Units do not result in changes in phenomenal experience (Marshall et al. 2018). The only difference Micro Units can make in phenomenal experience is the difference that leads to a change in

the state of the Macro Units they constitute. From this we can conclude: Macro Units exist intrinsically; Micro Units exist extrinsically. It follows that, for the complex, Macro Units are its *intrinsic constituents*, while Micro Units are only its *extrinsic constituents*. Finally, we can say: Macro Units do not emerge from Micro Units. As Micro Units do not exist intrinsically, for Macro Units to be emergent would mean to emerge from nothing. Instead, we can define the metaphysical relation between Macro Units and Micro Units as a relation of *subsumption* (from Latin "*subsume*", i.e., to take under).

First-order & Higher-order Mechanisms

The second case, the relation between first order mechanism and higher order mechanism, goes more smoothly: either a candidate mechanism exists (i.e. it is a mechanism) or not. To assess if a candidate mechanism exists one has to check Information, Integration and Exclusion (Oizumi et al. 2014). To assess if a mechanism is part of a system one has to check Intrinsicality, Composition, Information, Integration and Exclusion. We commonly label mechanism with letters, such as A, B, C, AB, ABC and so on, but these are in fact just labels. Higher order mechanisms have no relation of emergence with First order mechanism. Each mechanism, if it exists, it does so independently from the others. Therefore, we can simply say that mechanism of any order simply co-exists and do not emerge.

Cause-Effect Structure & Main Complex

Finally, does the Cause-Effect Structure emerge from the complex? Once again it is important to consider that this question is asked within the context of IIT and therefore it is meaningful only if answered following the theory's principles. We already saw how consciousness is equated with existence. To exist means to have experience, and an experience can be described as a Cause-Effect Structure (Tononi, 2015). However, consciousness is our starting (and only) point in our knowledge of the world. Hence, for IIT the problematic component of the Mind-Body Problem is not the "mind" (experience) but the "body" (the world). We know what the mind is, what we do not know (and cannot know directly) is the world. Hence the physical properties that we use to describe the complex are just a way to operationalize what we can observe and manipulate, and by doing so we unfold the Cause-Effect Structure of the complex. The Cause-Effect Structure is what the complex appears us to be. It

would be a mistake to consider that the Cause-Effect Structure is derivative in any meaningful way from the complex, let alone emergent. We commonly say that the complex *specifies* its Cause-Effect Structure, but it is important to notice that, ontologically speaking, the Cause-Effect Structure is primitive compared to the complex, which is perceived by an observer.

Conclusion

To conclude, in IIT, as much as there is no ontological reduction (Tononi, 2017), there is no ontological emergence. Since entities *are* (they exist), they cannot emerge from entities that *are not* – it would be a contradiction. Since existence is a central notion within IIT, it follows that nothing emerges, everything *is*. We can define entities in different terms, such as phenomenally or operationally (i.e. physical) but given the fact that experience is epistemologically fundamental, we consider entities from their intrinsic point of view and not from how they appear to us. As we stated in the beginning though, there is neither a fixed notion of emergence in the literature, nor evidence for convergence of positions. For this reason, while the most common definition of emergence do not currently fit in IIT, it could be possible in the future that some peculiar definition of emergence match the descriptions provided by the theory.

Bibliography

Anderson, P. (1972). More is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science. reprinted in Bedau, M., Humphreys, P. (eds.) (2008), *Emergence. Contemporary Readings in Philosophy and Science*, Cambridge: MIT Press.

Bunge, M. (2003). *Emergence and convergence*. Toronto: University of Toronto Press.

De Wolf, T., & Holvoet, T. (2005). Emergence Versus Self-Organization: Different Concepts but Promising When Combined, 1-15.

Ellis G. (2012) Top-down causation and emergence: some comments on mechanisms. *Interface Focus*. 2126–140.

Hooker, C. (2011). Conceptualizing Reduction, Emergence and Self-Organization in Complex Dynamical System, in C. Hooker (ed.), *Philosophy of Complex Systems*, Oxford: North Holland.

Humphreys, P. (2016). *Emergence. A Philosophical Account*. Oxford: Oxford University

Kauffman, S. A. (1993). *The origins of order*. New York: Oxford University Press, USA.

Marshall W, Albantakis L, Tononi G (2018). Black-boxing and cause-effect power. *PLoS Computational Biology* 14(4): e1006114.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, *10*(5).

Thurner, S., Corominas-Murtra, B., and Hanel, R. (2017). Three faces of entropy for complex systems: Information, thermodynamics, and the maximum entropy principle. *Physical Review E*, 96(3):032124.

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1).

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450-461.

Appendix B – Glossary

Axioms: Essential properties of any conceivable experience.

Background conditions: fixed external constrains on a set of elements in a state.

Candidate system: a set of elements which is taken under consideration. For the evaluation of integrated information.

Cause-Effect repertoire: the probability distribution of potential past and future states of a system that is specified by a mechanism in a state.

Complex: A set of elements in a state that generate a local maximum of integrated information Φ^{max}.

Concept: see distinction.

Conceptual Structure: see MICS.

Cut: see partition.

Distinction: links a maximally irreducible cause with a maximally irreducible effect within the cause-effect structure.

First order mechanism: A mechanism that includes only one element.

Higher order mechanism: a mechanism that includes more than one element.

Intrinsic Perspective: how the system is for itself, independently from any external observer.

Macro units: a unit composed by a set of micro-units

Main Complex: the physical substrate of consciousness

Mechanism: any proper or improper sub-set of elements that belong to the candidate system, which has a maximally irreducible causal role within the system. A mechanism can be first order (one element) or higher order (more than one element). In any case a mechanism is a unitary entity, whose causal role is not reducible to its components.

Micro units: A set of units that constitutes a macro unit.

MICS: Maximally irreducible cause-effect structure. The cause-effect structure of a complex in a state that corresponds to a maximum of integrated information Φ^{max}. It is identical to an experience, or quale (in the broad sense).

MIP: Minimum information partition. The partition of a distinction of a system that makes the least difference to its information.

NCC: Neural Correlates of Consciousness; the minimum neuronal mechanism jointly sufficient for a conscious experience. The full NCC are the minimum neuronal mechanism jointly sufficient for any conscious experience.

Partition: division of a set of elements into causally independent parts, technically performed by injecting noise in the edge between two or more nodes.

Postulates: operationalization in cause-effect power of the axioms. In other words, postulates are the condition of possibility for the axioms in physical terms.

Physical substrate of consciousness (PSC): the set of elements that specifies a cause-effect structure identical with an experience.

Purview: any set of elements in a candidate system over which the cause-and-effect repertoires of a mechanism in a state are calculated.

Q-Structure: see MICS

Q-Shape: a subset of a Q-Structure. Typically, a set of distinction bound by relations.

Quale: the qualitative feeling of phenomenal distinction within an experience.

TPM: Transition probability matrix. A matrix that specifies the probability with which any state of a set of elements transitions to any other state of the same set of elements.

The TPM is obtained either by combining the activation function of each element or by perturbing the candidate system in all its possible states.

Relation: maximum irreducible overlaps among the purviews of two or more distinctions.

Appendix C – Calculating PHI

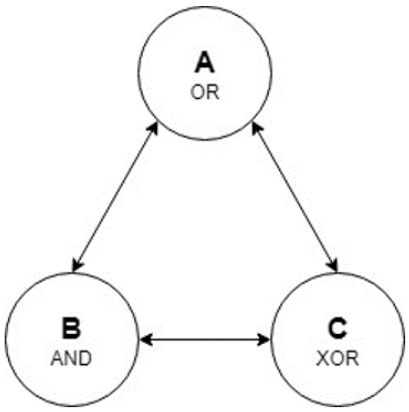
In this section I will discuss integrated information with more technical details, following supplementary materials in (Mayner et al., 2018). The example of the network ABC is due to (Oizumi et al., 2014). A complete discussion and presentation of the theory from a formal point of view can be found elsewhere (Albantakis et al., 2019; Barbosa et al., 2021; Barbosa et al., 2020; Gomez et al., 2021; Haun & Tononi, 2019; Mayner et al., 2018; Oizumi et al., 2014).

While IIT is a fully fledge theory, its formalism has been improved over the years with major changes between an update and the others. Those who are genuinely interested in IIT should always consider the formalism of the latest available version. When I started working on this dissertation, the most recent version of IIT was the so-called 3.0 version (Oizumi et al., 2014). Few months before I completed my work, some papers were published with the initial formalism of IIT 4.0 (Barbosa et al., 2021; Barbosa et al., 2020). However, the full version of IIT 4.0 is still not published. For this reason, I focused my analysis on the ontological and epistemological issues of the theory. However, I believe that a brief presentation of the general principles between the mathematical model of IIT could have been useful to give a primer to the reader. This appendix is meant for a reader who does not have a background in quantitative sciences but has interest in understanding IIT. In what follows, I show how to compute Φ and φ for a simple three-nodes network following the formalism of IIT 3.0 with the updated background conditions (Mayner et al., 2018; Oizumi et al., 2014). Clearly such a simple network cannot capture the complexity of a human brain with

billions of neurons – it is not its purpose either. Small "toy" systems are used to present clearly the formalism of the theory and its consistency (Albantakis, 2021).

Intrinsic Existence and Composition

To study a physical system, we can model it as a network of interconnected elements, each of which is in one of at least two states. The state of each element is described by its input-output function that determines the element transition from one state to another. For example, imagine a switch A connected by a non-noised edge to two inputs: A can be either on (1) or off (0) and it turns on in a fully deterministic way when at least one of its inputs is on: A can be modeled as a simple logic-gate with an OR activation function, hence, assuming that A will have two inputs, B and C, A's state can be fully characterized by its Transition Probability Matrix (TPM): if both inputs are OFF then A's state will be OFF and in the three remaining cases A's state will be ON. A similar TPM can be obtained for each node of the network, for example B's activation function is an AND gate and C's activation function is a XOR[51] gate. Combining the TPMs of each element we obtain the System TPM, which fully characterize the system behavior. Note that the edges (connections) between elements are deterministic, meaning that each node receives an input which is either 1 or 0, without any noise, though in general can be probabilistic.



T_0	T_{+1}
ABC	ABC
000	000
100	001
010	101
110	100
001	100
101	111
011	101
111	110

Figure 1: the network ABC and its TPM

[51] A XOR logic gate has an exclusive OR function: its truth table with two elements is: 00 = 0, 10 = 1, 01 = 1, 11 = 0.

Note that usually the activation function for an element is unknown to the experimenter, however the TPM of the system[52] can be obtained by observing and perturbing the elements in every possible state and then combining them:

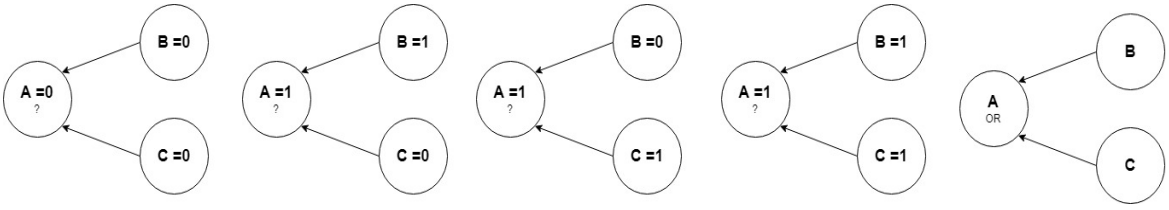


Figure 2: assessing A's activation function by perturbing the elements in every possible state.

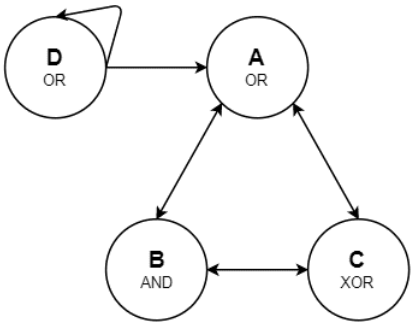
Once we see how the nodes change according to their inputs' perturbations, we can determine their activation functions. In this case, A turns ON if any of its inputs are ON, and therefore A is an OR gate. Likewise, knowing the activation function of a node without knowing the states of its inputs is *per se* informative. In fact, to know that A is an OR gate and A is ON at T_0 means that of the four possible states its input can be at T_{-1}, one (both inputs OFF) is excluded. Conversely, knowing that A is an OFF at T_0 is highly informative, as it means that of the four possible states its input can be at T_{-1}, three (any input ON) are excluded. Finally, note that ON and OFF are simply labels without any real meaning, they depend exclusively on how the model is built. Therefore, an OR and an AND gates are equivalent but flipped.

Background Conditions

The example above describes a simple case: the system under examination is a three nodes network completely isolated from external influences. Such a case does not reflect a real system. In fact, most likely one of the issues that the experimenter will face is how to separate the system under examination from its environment. For example, consider this extended version of the previous network, the network ABCD in its state 1000[53]:

[52] From now on, the states of the network ABC will be address both by their label (for example 000 to indicate all nodes are off) and by their number (#0 to indicate 000, #1 to indicate 100, #2 to indicate 010 and so on). Through all the present article states are enumerated according to the little-endian convention.

[53] Importantly, Integrated Information is state dependent. This means that integrated information is always calculated over a specific state of the system, not generically. In the rest of the section the state of the network is always A = 1, B = 0, C = 0, as explained in the text.



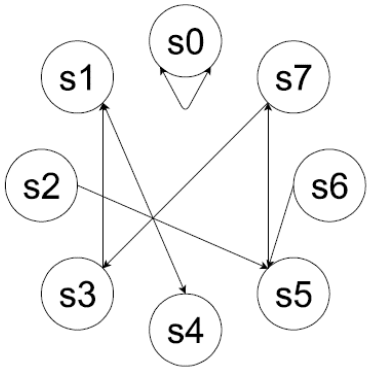
T_0	T_{+1}
0000	0000
1000	0010
0100	1010
1100	1000
0010	1000
1010	0110
0110	1010
1110	1100

T_0	T_{+1}
0001	1001
1001	1011
0101	1011
1101	1001
0011	1001
1011	1111
0111	1011
1111	1101

Figure 3: the network ABCD and its TPM

In this case, we still want to consider ABC as candidate system and treat D as a background condition, however we do not want to simply cut D out of the picture, as its causal links with the candidate system are still relevant, we need want to fix D in its current state. This is an advancement from IIT 3.0, where elements outside the candidate system were fixed in their previous state (Mayner et al. 2018).

To do so, we condition ABC on D with D = 0, by considering only the relevant TPM (in other words, in the above picture we can ignore the second column as it includes only states with D = 1). Then, future states of D are ignored by marginalizing out D. To do so, we develop the TPM into probability distributions for each state, note that since the network is deterministic (each state at T_{+1} has probability either 1 or 0) this means that, for example, state '0000' at T_0 has probability 1 to lead to state '0000' at T_{+1} and probability 0 for every other state. Now, to marginalize out D we sum (and then normalize according to Bayes rule) the probability of all states that differ only for D. For example, the first state of each column is identical for the first three elements but differs for D (0000 and 0001). Then, we sum the probability of both states at T_{+1} (0+1) and we have the probability for state 000 (since D has been marginalized out) at T_{+1} for 000 at T_0. We repeat the same operation for each state, and we obtain the TPM presented in the previous section, except that this time we know that D is present but treated as background condition. We can now expand the TPM to highlight that every row represents a probability distribution for a given state (rows) at T_0 to lead to another state (columns) at T_{+1}:



		#0	#1	#2	#3	#4	#5	#6	#7
		000	**100**	**010**	**110**	**001**	**101**	**011**	**111**
#0	**000**	1	0	0	0	0	0	0	0
#1	**100**	0	0	0	0	1	0	0	0
#2	**010**	0	0	0	0	0	1	0	0
#3	**110**	0	1	0	0	0	0	0	0
#4	**001**	0	1	0	0	0	0	0	0
#5	**101**	0	0	0	0	0	0	0	1
#6	**011**	0	0	0	0	0	1	0	0
#7	**111**	0	0	0	1	0	0	0	0

Figure 4: ABC dynamics state transitions in its State-Space are captured by its TPM.

The TPM captures the dynamics of the candidate system or, in other words, it describes entirely the system behavior. However, given the premise that there is a double dissociation between consciousness and behavior (behavior is not necessary for consciousness, and consciousness is not necessary for behavior) we argue that the TPM is not enough. So instead of focusing on the extrinsic behavior we want to assess the intrinsic causal dynamics of the system, namely how the present state constrains the past and the future states of the system. From now on, unless differently specified, the system will be considered in its state 100. Note that the causal analysis of IIT is state dependent and therefore the value found can vary significantly in different states.

Information: Cause-Effect Repertoires

Once the candidate set in a state is fully described by its TPM with fixed background conditions, the proper causal analysis can be applied to the system in order to determine how its current state is constrained by past and future states. In this case all possible non-empty subsets of the candidate system are called candidate mechanism, and their respective causal properties are assessed: this process is known as unfolding the system[54].

[54] Unfolding is the technical name of the procedure that leads to the cause-effect structure starting with a network. It should not be confounded with the unfolding argument, see chapter 5.

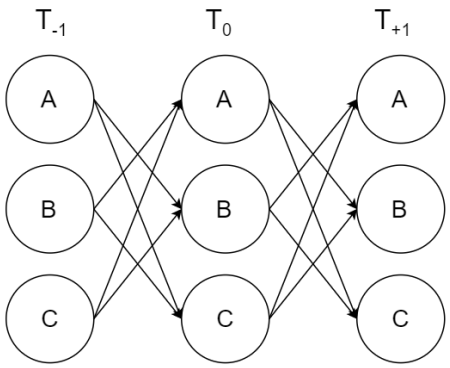


Figure 5: the unfolded diagram of the candidate system ABC.

For example, candidate system ABC presents the following candidate mechanism: [A], [B], [C], [AB], [AC], [BC], [ABC]. In this case we will refer as first order mechanism for mechanism [A], [B], [C], while we will refer as second order mechanism for [AB], [AC], [BC], and third order mechanism for [ABC] [55]. Once the candidate mechanism are individuated their causal properties are described in a formal way as probability distributions by specifying how much the mechanism in its present state constrains the system past and future states. Said constrains are formalized as Cause Repertoire and Effect Repertoire of the candidate mechanism, namely they are their probability distributions over states respectively at T_{-1} and T_{+1}.

Effect Repertoires

For the candidate mechanism ABC in its state 100, looking at the TPM we know that at the subsequent timestep the probability of state 001 is 1 and the probability of all other seven states is 0. However, we want to know more, we want to know how every candidate mechanism is responsible for what effect. To do so, we want to evaluate the effect of the candidate mechanism over every possible (non-empty) subset of elements at T_{+1}, these subsets are called *purviews* of the candidate mechanism.

Imagine for example we want to evaluate the effect of candidate mechanism C over its purview BC. To do so we need to fix C in its current state at T_0 and then perturb A and B into all possible states, with equal likelihood, observing their effect on purview BC at T_{+1}. However, since B and C at T_{+1} have common inputs at T_0 (they both receive inputs from A) we need to introduce *virtual elements*. Virtual elements are

[55] Non-first order mechanism are also known collectively as 'higher order mechanism'.

elements at the previous timestep which send inputs individually to one node: in this case we will introduce at T_0 the virtual element A_b (which sends its input only to B at T_{+1}) and the virtual element A_c (which sends its input only to C at T_{+1}). This will generate a new TPM for the system which is called virtual TPM as it does not describe 'actual' transition probabilities but only virtual ones. Now we can finally perturb A_c A_b and B in all possible states and assess their effect over BC. To do so, in the Virtual TPM we marginalize A out at T_{+1} (as it is not part of the purview). Then we marginalize out A_c, A_b and B at T_0, and finally we consider the current state of C (C being equal to 0 in state 100) and we obtain a probability distribution:

	BC = 00	BC = 10	BC = 01	BC = 11
C = 0	0.5	0	0	0.5

For C = 0 at T_0, BC at T_{+1} will have probability 0.5 to be either 00 or 11 and probability 0 to be either 01 or 10.

Finally, we can expand this result to the whole state space, obtaining a distribution over all ABC states. To do so we simply multiply this distribution by the unconstrained distribution over non purview elements. In this case only A is not part of the purview, and being an OR gate its unconstrained distribution is:

A = 0	A = 1
0.25	0.75

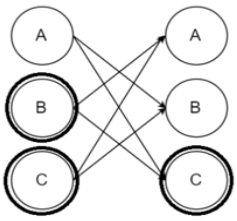
The tensor product between these two distributions gives as output the Effect Repertoire of Candidate Mechanism C over its Purview BC.

Cause Repertoire

Now we want to assess how each mechanism in a state independently constrains the past state of the system. In this case, a purview is a (non-empty) subset of elements at the previous timestep. Once again, in the case of higher order mechanism we will use virtual elements to perturb independently the inputs of each elements of the candidate mechanism[56]. Contrary to the effect repertoire, a remarkable difference is that in the Cause Repertoire the Unconstrained Distribution is the uniform distribution: no previous state is *a priori* more relevant than the others. Moreover, due to conditionally independence (a condition assumed to be true in order to rule out instant causation) the Cause Repertoire of higher order mechanism is given simply by the tensor product

[56] For example, if we consider candidate mechanism C over its purview BC we need to introduce virtual element A.

of the repertoires of individual first order mechanism. Consider for example candidate mechanism C over its purview BC.



T_{-1}	T_0							
	000	**100**	**010**	**110**	**001**	**101**	**011**	**111**
000	1	0	0	0	0	0	0	0
100	0	0	0	0	1	0	0	0
010	0	0	0	0	0	1	0	0
110	0	1	0	0	0	0	0	0
001	0	1	0	0	0	0	0	0
101	0	0	0	0	0	0	0	1
011	0	0	0	0	0	1	0	0
111	0	0	0	1	0	0	0	0

BC	**C**
00	0.5
10	0.5
01	0.5
11	0.5

Figure 6: Candidate Mechanism C and its Purview over BC. The full TPM and the Cause Repertoire of C over BC in state 100.

Note that this TPM goes from T-1 to T0 instead of T0 to T+1; therefore, in order to marginalize out A (as it is not included in the purview) we need to sum and normalize rows instead of columns. Then, in order to marginalize out A and B (as they are not part of the candidate mechanism) we need to sum columns instead of rows. Finally, we consider C in its present state (0, as in 100) and we obtain the Cause Repertoire of C over BC in state 100.

Integration

The Cause-Effect Repertoire captures the selectivity of causes and effects of the mechanism over the system. However, if a candidate mechanism has no new power in terms of causes and effects, i.e., if the mechanism as a whole does not add any causal interaction within the system, then the mechanism is reducible. For example, if candidate mechanism XY does not do anything more than X and Y then there is no point in considering X and Y together a mechanism in first place. A difference that does not make a difference is no difference.

In order to assess if the repertoire of a mechanism is reducible, we need to 'cut' its purview in two parts: for example, we can cut C out of the purview when we assess candidate mechanism AC over its purview ABC. Then the effect of the candidate mechanism AC is assessed over purview AB and the unconstrained repertoire of C: $\frac{AC}{ABC}$ is partitioned into $\frac{AC}{AB}$ and $\frac{\emptyset}{C}$.

We then proceed to calculate the effect of AC over AB

	AB = 00	AB = 10	AB = 01	AB = 11
AC= 10	0.5	0.5	0	0

And the unconstrained distribution of C:

C = 0	C = 1
0.5	0.5

And finally, their tensor product (to expand to the full space):

	000	100	010	110	001	101	011	111
AC= 10	0.25	0.25	0	0	0.25	0.25	0	0

Comparing the obtain Effect Repertoire with the original Effect Repertoire of AC over ABC we notice that the two are identical. Therefore, there is no gain in including C in the Purview of mechanism AC.

However, $\frac{AC}{ABC}$ can be partitioned in multiple ways:

$$\frac{\emptyset}{A} \times \frac{AC}{BC} \; ; \; \frac{\emptyset}{B} \times \frac{AC}{AC} \; ; \; \frac{\emptyset}{AB} \times \frac{AC}{C} \; ; \; \frac{\emptyset}{C} \times \frac{AC}{AB} \; ; \; \frac{\emptyset}{AC} \times \frac{AC}{B} \; ;$$

$$\frac{\emptyset}{BC} \times \frac{AC}{A} \; ; \; \frac{\emptyset}{ABC} \times \frac{AC}{\emptyset} \; ; \; \frac{\emptyset}{A} \times \frac{C}{ABC} \; ; \; \frac{A}{A} \times \frac{C}{BC} \; ; \; \frac{A}{B} \times \frac{C}{AC} \; ;$$

$$\frac{A}{AB} \times \frac{C}{C} \; ; \; \frac{A}{C} \times \frac{C}{CB} \; ; \; \frac{A}{AC} \times \frac{C}{B} \; ; \; \frac{A}{BC} \times \frac{C}{A} \; ; \; \frac{A}{ABC} \times \frac{C}{\emptyset}$$

For each possible partition is calculated the correspondent repertoire; then is measured the Earth Mover's Distance (EMD) between these newly obtained repertoires and the original repertoire. The EMD is a distance which quantify the 'cost' of transferring the minimum amount of earth to equate two different distributions of earth. This cost is given by the amount of earth moved time the distance it travels.

Among all partitions, the one with the minimal distance to the original repertoire constitute the Minimum Information Partition (MIP), the partition which lead to the minimum loss of information. Therefore, the distance (EMD) between the unpartitioned repertoire and the MIP amount to the irreducibility of the unpartitioned repertoire. This quantity is called 'integrated information' and it is captured by φ (small phi). In other words, φ measures the information present in the mechanism as an integrated entity and its irreducibility.

Exclusion: Maximally Irreducible Cause-Effect Repertoire

Given a candidate mechanism (for example the higher-order mechanism AB) we find its Cause-Effect Repertoire over all of its possible Purviews.

Then, for each of these we find the correspondent MIP and its φ. To find the Maximally Irreducible Cause of the candidate mechanism we select its Cause Repertoire with the highest φ value (φ^{max}). To find the Maximally Irreducible Effect of the candidate mechanism we select the Effect Repertoire with highest value of φ (φ^{max}).

The Maximally Irreducible Cause-Effect Repertoire of the candidate mechanism AB is given by its φ_{cause} and φ_{effect} values, which together constitute the Distinction[57] AB specified by mechanism AB.

The irreducibility of the mechanism as a whole (φ) is given by the minimum value of its Maximally Irreducible Cause and Maximally Irreducible Effect. It follows that if either the cause or the effect of a mechanism is reducible, i.e., its φ is 0, then the mechanism has a whole has $\varphi = 0$ and therefore is reducible, if a mechanism is reducible then it does not specify a Distinction.

The set of all irreducible Distinction specified by every system mechanism constitute the system's Cause-Effect Structure.

Systems of Mechanism

In the previous sections we have presented the tools to assess the irreducibility of a mechanism and quantify the value of its integrated information. Now instead of

[57] 'Distinction' is the term currently used by IIT theoreticians. In previous versions of IIT (such as IIT 3.0) it is possible to find the alternative 'Concepts'.

mechanism we look at systems of mechanism to assess whether of or not a collection of mechanism constitutes a genuine set or is just an arbitrary collection. Once again, the key is to assess the value of integrated information at the System level. To do so we will 'cut' the system in two. Formally, a unilateral cut does not remove edges between nodes altogether but inject noise in the node's output instead.

Consider for example the system cut A → BC over the system ABC means that B and C inputs from A will become noisy and therefore A will provide to both independently random input, while other connections will remain intact.

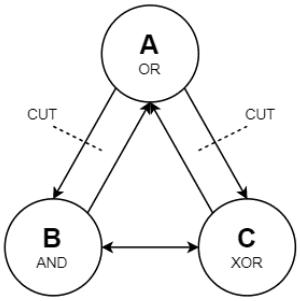


Figure 7: The Inputs from A to B and C are injected with noise.

Once again, we have to calculate the TPM individually for each mechanism and then combine them in order to obtain the new TPM of the system cut. Once each mechanism TPM is obtained and expanded to the full space states their tensor product will provide the system cut TPM.

	000	**100**	**010**	**110**	**001**	**101**	**011**	**111**
000	0.5	0	0	0	0.5	0	0	0
100	0.5	0	0	0	0.5	0	0	0
010	0	0.5	0	0	0	0.5	0	0
110	0	0.5	0	0	0	0.5	0	0
001	0	0.25	0	0.25	0	0.25	0	0.25
101	0	0.25	0	0.25	0	0.25	0	0.25
011	0	0.25	0	0.25	0	0.25	0	0.25
111	0	0.25	0	0.25	0	0.25	0	0.25

With the new TPM we can re-calculate the Cause-Effect Structure and compare it with the original one: if the two Cause-Effect Structures are identical then the cut made no difference to the candidate system and therefore the candidate system was

not a system. On the other hand, if there is a difference, we can quantify that difference by measuring the distance (with EMD) between the two conceptual structures. In this case the EMD will measure the cost of moving the φ value of each Distinction to another Distinction; this cost is also called 'concept distance'.

For example, we can appreciate how the Cause-Effect structure of ABC changes when we do the unilateral cut A → BC. In this case, all Distinctions but one disappears: the cut made a difference.

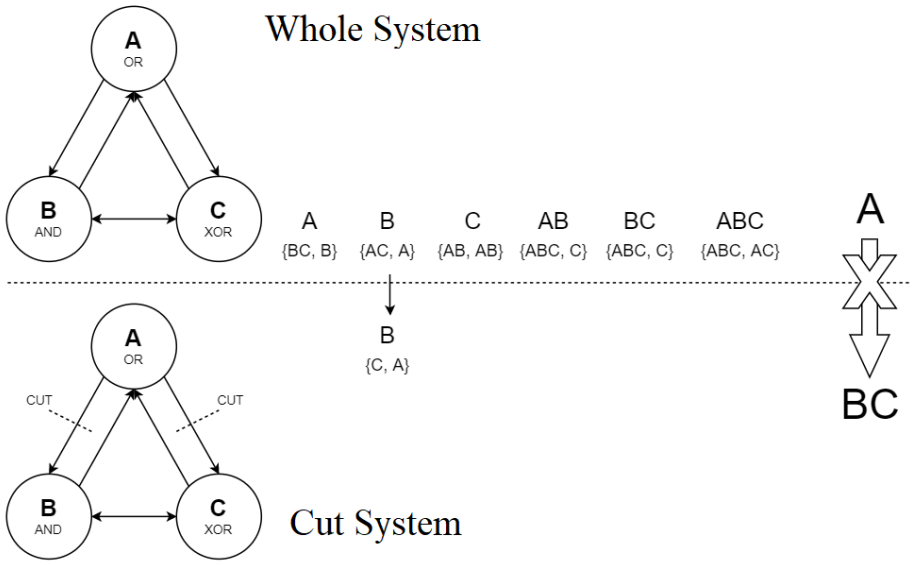


Figure 8: How the unilateral cut A → BC affects the Cause-Effect Structure of the system (represented by Distinctions and their Cause-Effect Repertoire). Only Distinction [B] 'survives' the cut.

We can now measure the distance between the original Cause-Effect Structure and the Cause-Effect Structure of the cut system. Its concept distance is given by the sum of EMD between their Cause Repertoire and the EMD between their Effect Repertoire. Since the new Cause-Effect Structure does not have a correspondent distinction for each of the vanished distinctions we need to compare their distance to the null concept: the null concept is not specified by any mechanism and its $\varphi = 0$. The null concept Cause-Effect Repertoire is the unconstrained Cause-Effect Repertoire.

Once we have accounted for each distinction, we sum all the distances in order to obtain an Extended Earth Mover's Distance (Extended EMD) value of distance between the original Cause-Effect Structure and the cut one. This quantity is called Integrated Conceptual Information and it is noted by Φ (Big Phi). The Integrated

Conceptual Information measures the irreducibility of the system given a certain system cut. Once again it is worth to point out that formally, if a cut does not make a difference to the system, then the distance between the original Cause-Effect Structure and the cut one is zero, and therefore $\Phi = 0$.

The following step consist in calculating Φ for each possible system cut. For example, given ABC there are 6 possible unilateral cuts: A → BC, B → AC, C → AB, BC → A, AC → B and AB → C. Generally speaking, for a system of n elements there are 2^n -2 possible System Cuts (the powerset except the whole set and the empty set).

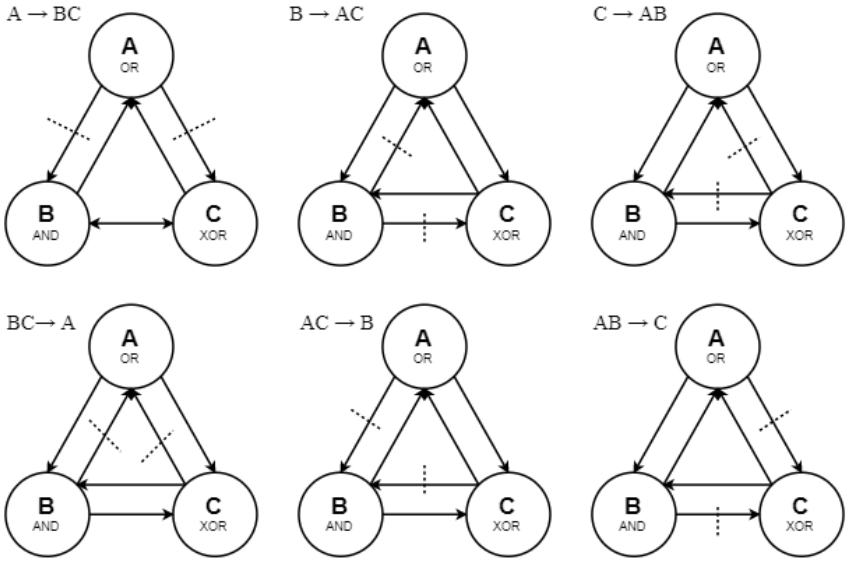


Figure 9: ABC is cut in all possible ways.

Among all cuts, the one which yields the lowest Φ-value is called Minimum Information Partition (MIP) and its Φ^{MIP} is the Φ of the whole system. Again, notice that due to the procedure to calculate Φ is impossible to obtain a negative number, therefore Φ is either equal or strictly greater than zero. For this reason, if there is a cut that makes no difference to the system then the system's Φ^{MIP} equals 0 and the system is reducible (it has $\Phi = 0$) meaning that is not conscious and does not exists as a system.

Finally, Φ has to be evaluated across multiple scales and for any subset and superset of the candidate system, or any system that partially (or completely) overlap with the candidate system. The system with the highest Φ-value, i.e., Φ^{MAX}, is called 'complex'. According to IIT, only a complex exists intrinsically as a subjective entity, thus defying clearly the borders (or causal borders) of the Physical Substrate of Consciousness (in

animals with brain, including Humans and Mammals, Neural correlates of Consciousness, NCC).

For example, consider the initial example with the whole network ABCD. In this case the possible subsystem that can be applied the IIT formalism are: [D], [AB], [AC], [AD], [BC], [CD], [ABC], [ABD], [ACD], [BCD], [ABCD].[58]

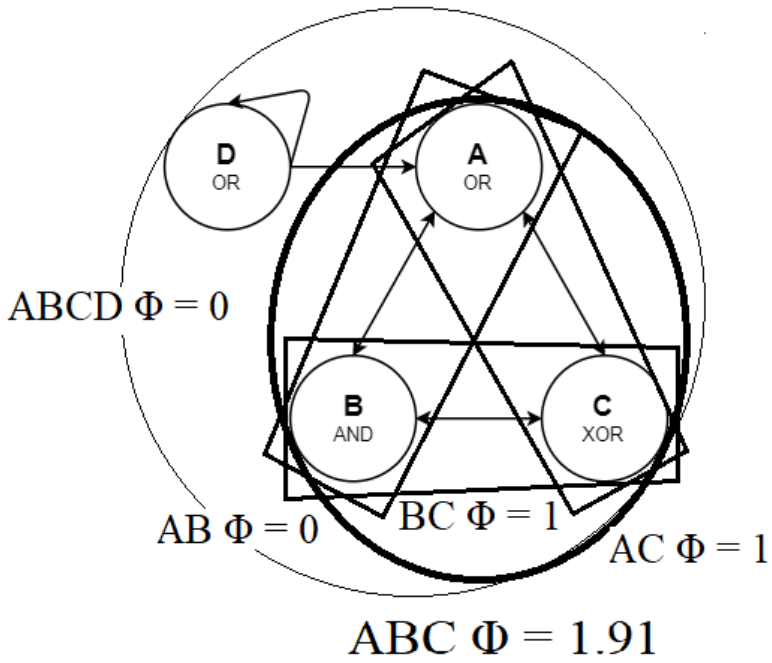


Figure 10: Φ is evaluated across any subset and superset of the candidate system in state 1000.

In this case it is self-evident that any system which includes D will have $\Phi = 0$, as there is a system cut D → ABC that makes no difference to the system, and any subsystem that includes D will have a similar cut. Generally speaking, feed forward system will present a cut that leads trivially to $\Phi = 0$, in fact, Integrated Information seems to require a certain number of feedbacks. ABC has a $\Phi^{MAX} = 1.91$. Therefore, according to the postulate of exclusion, only the system ABC exists intrinsically as a single entity (while A, B, C exist as parts of that entity).

[58] Note that these are systems, not mechanism or distinctions. [D] is included because given its reflexive edge (or self-loop) it has cause-effect power upon itself, formally this is rendered by the fact that [D] as a system can be analyze with the present formalism. Other elements have no reflexive edge therefore can exist only as mechanism.