MÉMOIRE DE MASTER 2

Philosophie des sciences, de la connaissance et de l'esprit

The Problem of Artificial Qualia

Auteur:

Adrien Wael BASILLE

Superviseur:

M. Pascal LUDWIG

2020-2021

Table of contents

Introduction

Is it possible to build a conscious machine? By "conscious machine", I mean an artifact that has qualitative experiences such as feeling pain, seeing the redness of a flower or enjoying the taste of coffee. What makes such experiences conscious is their *phenomenal character*: it is like something to have such experiences. As Thomas Nagel puts it, "*the fact that an organism has conscious experience at all means, basically, that there is something it is like to be that organism"* (Nagel 1974). There is something it is like to be me and I know this because right now I have phenomenal experiences, like that of a visual field full of colors. Presumably, there is also something it is like to be a bat perceiving its environment through echolocation, although a bat experience probably has a very different phenomenal character from that of humans. By contrast, most of us have the strong intuition that there is nothing it is like to be a table, a thermostat, a laptop or a robot. How to explain this difference? A straightforward answer is that animals have brains whereas artifacts don't. According to materialism[1], consciousness arises from brain activity. Although a brain is undoubtedly an immensely complex organ, it is nothing more than an organic machine for a materialist. From this perspective, it is reasonable to believe that a complete scientific understanding of the brain would allow us to replicate what it does to make consciousness arise in an artifact. If we managed to replicate brain processes in a robot, why not believe that it could become conscious? Here is for example what the materialist Daniel Dennett wrote on that matter:

> *Thinking in terms of robots is a useful exercise, since it removes the excuse that we don't yet know enough about brains to say just what is going on that might be relevant, permitting a sort of woolly romanticism about the mysterious powers of brains to cloud our judgment. If materialism is true, it should be possible ("in principle!") to build a material thing–call it a robot brain–that does what a brain does, and hence instantiates the same theory of experience that we do.* (Dennett 2006)

In this work, I will do exactly what Dennett recommends: thinking in terms of robots as a *useful exercise* to address a difficult philosophical question. This question, in a nutshell, is the following: is it in principle possible to have a complete theory of experience? And a necessary condition for a theory of experience to count as complete is that it should be able to

[1] I will use "materialism" and "physicalism" interchangeably in this work.

tell us how to build an experiencing machine. The underlying idea is that "*what I cannot create, I do not understand*" (Feynman 1988). In other words, it would be legitimate to claim that we understand consciousness only if we were able to create it ourselves.

In contemporary philosophy of mind, the question of the qualitative aspect of conscious experiences is often addressed in terms of *qualia* (Tye 2018). In a pre-theoretical and intuitive sense, qualia refer to the phenomenal character or "*what-it's-like-ness*" of a conscious experience: the painfulness of a pain experience or the redness of a visual experience of a red object. What I am wondering is whether or not qualia could be realized by artifacts. This is what I call *the problem of artificial qualia* and my first goal will be to explain what the problem is. The term "qualia" is used in different ways by different authors, sometimes causing confusion. Since it has a relatively long history, the concept of qualia is not free from implicit theoretical assumptions coming from the philosophical tradition in which it was born (Crane 2000).

A common way of defining qualia is to say that they are *phenomenal properties* of experience or of subjects. This is not how I will define them. Instead, I will use the term qualia as an abbreviation of "qualitative experience" and I will explain why. In the first part, I will define qualia by relying on a thought experiment. The experience of pain will appear as the typical example of what a quale is. If we had a complete theory of pain in the sense that we could create an artificial experience of pain, we would have solved the problem of qualia. As Hilary Putnam wrote, pain is of primary interest for the philosopher of mind:

> *The typical concerns of the Philosopher of Mind might be represented by three questions: (1) How do we know that other people have pains? (2) Are pains brain states? (3) What is the analysis of the concept pain?* (Putnam 1967)

Once equipped with a definition of qualia, I will discuss several points concerning the traditional way of conceiving them. I will try to understand why certain philosophers such as Dennett deny that qualia exist. What they deny is in fact the existence of phenomenal properties and it is legitimate to claim that this notion is problematic. Then, I will argue that the problem of artificial qualia is essentially a philosophical problem that cannot be solved by science alone.

The second part will be dedicated to an analysis of the concept of qualia. I will start with the problem of perception and present what is arguably the most popular theory of perception in cognitive science: representationalism. My aim will be to account for qualia in a representationalist framework. I will argue that qualia should be interpreted as perceptual modes of presentation. This will lead me to the adverbial analysis of experience. Adverbialism will appear as an appropriate way to account for qualia as I defined them. I will then question the possibility of modeling qualia. It will be shown that there is an important obstacle to this project: it is very difficult to distinguish between the structure of inner experience from that of outer reality. Thus the idea that experience is an "inner" process may be misleading.

It is in the third and last part of this work that the problem of artificial qualia will be tackled. What are the processes of the natural world that give rise to qualia? I will discuss computationalism and functionalism, two very close interpretations of the mind that are at the roots of contemporary cognitive science and philosophy of mind. There is undeniably an important sense in which *thinking* is a computational process. However, I will argue that computationalism is misleading when it comes to qualitative experiences. The idea that qualia are arising from computations happening in the head will be challenged. The role of the body and its interaction with the environment may also be important to understand experience. Then I will consider the claim that only *living* creatures can realize qualia. Finally, I will examine the possibility that qualia are fundamental processes of the universe.

I will rely on extensive literature mainly coming from twentieth and twenty-first century philosophy of mind. Taken individually, almost none of the ideas and arguments presented in this work are new. I believe the original aspect of this work lies in the way the ideas are presented and linked together in order to aim at the over-ambitious goal of designing conscious artifacts. As I explained, artificial consciousness is essentially a pretext to tackle what I believe is the most central problem of philosophy of mind. I will regularly come back to the example of machines and use it as a methodological tool to progress on the understanding of qualia.

1. The problem of qualia

1.1. Mary in the room with no pain

Mary's Room is a famous philosophical thought experiment proposed by Frank Jackson (1982, 1986, 2004) and abundantly discussed in philosophy of mind. Mary is the name of an imaginary scientist who has never seen any color because she is confined in a black and white room. She spent all her life reading books about color vision to the point that she knows everything there is to know about the physics, chemistry and neurophysiology of colors. Jackson used this thought experiment as the basis of what is called the *"Knowledge argument"* (Jackson 1982). When Mary leaves her black and white room, so the argument goes, she learns something new about the world by looking at a red tomato for the first time. Since Mary was supposed to have acquired "*all the physical information there is to obtain about what goes on when she sees ripe tomatoes*" (ibid.), the conclusion of the argument is that there exists non-physical information. My aim in this first part is not to discuss the Knowledge argument but rather to use a variant of *Mary's Room* thought experiment to outline a definition of qualia that will serve as a starting point for this work.

Let's suppose that instead of being deprived of the subjective experience of colors, Mary undergoes a treatment making her insensitive to pain. Every day since birth, she is forced to ingest a very strong analgesic that completely eliminates her body aches and pains. In other words, Mary never experienced pain in her life, she does not know what pain feels like. As in the original thought experiment, Mary receives a scientific education, except this time she specializes in the neurophysiology of pain and becomes a super-expert in the domain, knowing everything there is to know about what happens in the body and brain of a human being when hurted. Moreover, a chip was placed in Mary's brain in order to retrieve all the nerve signals coming from her nociceptors. The chip then transmits the signals to a pair of connected glasses equipped with transparent screens that Mary wears permanently, informing her visually of her own body damages in real time. For example, when she puts her hand on a hot plate, the exact parts of the hand being burned appears in her visual field, together with a very complex color code that informs her of the exact type of burn her hand is exposed to. The visual representation of her body damage that she can access through her glass is sophisticated enough for her to be able to distinguish between the different types of damages,

at least as much as what a typical human can distinguish when feeling pain. In these conditions, there is a sense in which Mary is *conscious* of her impaired body tissues. She is conscious in the sense that she is *aware* that her hand is burning and she can use the precise information conveyed by her glasses to infer everything she could have inferred if she felt pain. There is nothing she cannot *know* about the properties of her body that she could have known if she felt pain. Here, I am talking about *factual* knowledge, and I will get back to this in part 1.2. What I mean is that there is nothing true she cannot say about the state of her body that she could have said if she had been able to feel pain. She knows all the bad consequences that the damage can cause, since she is an expert in biology. By hypothesis, she knows way more about all the properties of her body than any normal woman can know about hers when feeling pain. Let's now imagine that one day, her treatment is stopped. She puts her hand on a hot plate and, for the first time of her life, she feels her hand burning. She tells herself "*So that's what it's like to be in pain! What a terrible feeling, now I get why people hate it so much!*".

Although it involves a complex technical device that is arguably not easily deployable in practice, this variant of Jackson's original scenario is less prone to some typical objections. In particular, Dennett (2004) argued against the traditional understanding of *Mary's Room* original experiment by claiming that, not only Mary wouldn't discover anything new when leaving the black and white room, but she would even be able to recognize that a blue banana is unnatural. According to Dennett, if we accept that Mary really knows *everything* that could be known about the physical causes and effects of color vision, then we must accept that she knows what thoughts she would have in front of a yellow object. Therefore, he argues that Mary is able to deduce that a blue banana is blue as opposed to yellow in virtue of the effect that the blue banana has on her cognitive system. In other words, she would make the correct color distinctions because by hypothesis she knows all the physical information. I agree with Dennett that it is very hard to imagine what acquiring all the physical information about color vision really entails. However, this kind of objection is greatly weakened in the case of pain. The reason is that pain has an *intrinsic value* that is very hard to deny: it *feels* bad, unpleasant (Goldstein 1989). Of course, while under her painkiller treatment, Mary can know *as a fact* that pain is an unpleasant feeling that people usually try to avoid. Furthermore, she may know what an unpleasant experience is, for instance because she already experienced an unpleasant smell. That said, it is difficult to maintain that Mary is capable of inferring the particular quality of a pain experience merely from her book knowledge and her other unpleasant

experiences. The main reason is that a qualitative experience such as pain is not something that can be inferred at all. It is also true for color experiences but the qualitative aspect of pain experiences is arguably more salient. When it comes to colors, there is an ambiguity between the quality of the object itself (the redness of the ripe tomato) and the subjectively perceived quality (phenomenal redness). I will come back to visual perception in more detail in the second part. I prefer to start with pain because pain has the advantage of being unambiguously and essentially characterized by the way it is experienced. In fact, the word pain *denotes* the qualitative experience itself. This is reflected in the way the *International Association for the Study of Pain* defines pain: "*an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage*" (Raja & al. 2020).

Mary's super-glasses thought experiment is meant to show that one's conscious access to her own body damage does not necessitate pain. During her treatment, Mary is conscious of her body states thanks to her connected glasses. She is consciously informed that her hand is burning but there is nothing unpleasant about this conscious experience. Of course, pain is arguably a more efficient way of being informed about body damage than visual stimuli coming from the super-glasses. The unpleasantness of a pain experience is most likely a product of natural selection: those who felt bad when hurted avoided damaging their body and thus increased their chance of survival. I will get back to the evolutionary explanations of feelings in part 3.4. For now, the sole purpose of the thought experiment is to point out that pain can be interpreted as a particular *way of consciously experiencing* one's own body damage. This is important in the context of a discussion about artificial consciousness. For an artifact to feel pain, it is not sufficient that it becomes aware of its own physical impairment. Indeed, a conscious robot could experience its body states visually, in the same way as Mary does with her super-glasses. Pain refers to the way a normal experience of body damage is, to *what it is like* to experience it. Since it is like something to experience pain, pain is a qualitative experience. It is in this precise sense that a pain experience is a *quale*. When Mary feels pain for the first time, I want to say that she undergoes an experience that she never had, and I want to call this new experience a *quale* of pain. In other words, I am using the term quale as a *synonym* of "qualitative experience". A pain experience *is* a quale of pain. A quale is something that happens. As it will become clearer in what follows, this is an important point because it is not how the term is traditionally used. I will come back to this point in part 1.5. For now, it is just a matter of *definition*: if it is like something for a subject to undergo an

experience then this experience is qualitative. And a qualitative experience is a quale, it is just an abbreviation.

We can say that the first time Mary feels pain, she discovers a quale of pain. Now, Jackson's original *Knowledge argument* was meant to show that Mary acquires new *information* when discovering what pain is like. What is the relationship between qualia and information?

1.2. Qualia and phenomenal information

Does Mary acquire knowledge when she experiences pain for the first time? It is tempting to answer in the affirmative by claiming that Mary learns what it's like to feel pain. Although she tells herself "*this is what pain feels like*", where "*this*" refers to the quale of pain, it would be controversial to claim that this constitutes a *fact*. A fact is by definition something that can be true or false. "*This is what pain feels like*" can hardly be interpreted as a fact according to this definition. David Lewis argued that what she learns is not a new fact but a new *ability*. According to the "*ability hypothesis*", feeling pain is not a matter of acquiring information ("*know-that*") but should rather be understood as an ability ("*know-how*") (Lewis 1988). The ability hypothesis is compatible with the definition of qualia of the previous part. Indeed, pain is a particular way of consciously accessing information about one's own body damage. By taking an analgesic, Mary's capacity to access her injuries *via* pain is inhibited. However, thanks to her connected glasses, she can be aware of her body states in an unusual way. If Mary continues to wear the glasses after stopping the treatment, she accesses the same factual information about her own burning hand in two different ways. She has two distinct *abilities* serving one and the same purpose: to acquire information about her body tissues. Lewis opposes the ability hypothesis to the "*hypothesis of phenomenal information*" (ibid.), according to which there exists information *about the experience* itself, *in addition* to the physical information accessed through experience. Does it make sense to talk about phenomenal information? To answer this question, a clarification of the notion of information in this context is needed.

In a straightforward sense, which is the one Lewis seems to have in mind, the word "information" refers to *factual* information and can be used interchangeably with "knowledge". In this respect, the claim that Mary acquires new information when she feels pain is tantamount to a defense of the Knowledge argument. As just explained, the main flaw

of this argument is that there is nothing factual Mary did not know about pain before experiencing it subjectively. This is because, by hypothesis, everything factual was written in her scientific books. Nevertheless, there is an intuitive sense of the term "information" according to which the quale of pain can be interpreted as phenomenal information without contradicting Lewis' ability hypothesis. The notion of phenomenal information is useful because it highlights the fact that the painfulness of Mary's experience is meant to *inform* her of her body damages. In this view, the quale of pain is suitably interpreted as an information-bearing *signal*. In common language, the term "information" is ambiguous because it can equivocally refer to the signal itself or to what is signaled. For example, when an indicator light on the dashboard of a car signals that it is almost out of gas, it is common to say that the light itself constitutes information. In Mary's case, her physical injuries were signaled by a visual experience. Now that her treatment is stopped, her injuries are signaled by a pain experience. Both qualitative experiences are signals of different *forms*. It is in virtue of their particular forms that Mary is *informed* of her body states. Qualia constitute phenomenal information or "*phenomenally realized information*" (Chalmers 1996, p. 266) in the sense that they are the *forms* that an experience can take.

Thus it is important to insist again that "phenomenal information" has a different meaning than the one Lewis argues against, since it does not refer to information *about experience*. Lewis refuted the phenomenal information thesis because, as a physicalist, he found it unacceptable that Mary could learn something new when feeling pain for the first time. According to the definition of phenomenal information that I outline here, it would be a mistake to claim that Mary can *learn* it. To say that the quale of pain constitutes phenomenal information means that experience is an *informational process*, the very process that makes it possible to learn anything. To experience the world in a certain way is an ability, namely the ability to consciously access information about things in the world. If Mary had neither the ability to feel pain nor her super-glasses, she would have no way to *consciously* access information about her injuries. The notion of phenomenal information emphasizes the fact that experiences are informational in virtue of being qualitative. Another way to put it is to say that a conscious experience has its informational content in virtue of its phenomenal character. The quale of pain carries information about body damage in virtue of its unpleasant quality. This is why, in the real world, people that do not have the ability to feel pain are in great danger. People who suffer from congenital insensitivity to pain usually die young because they are not aware of their injuries or illnesses (Linton 2005). This is because they

are not *informed* of their injuries, there is nothing that *signals* it. It is in this precise sense that it is meaningful to say that qualia constitute phenomenal information, and this is not incompatible with the ability hypothesis.

To say that the quale of pain usually informs about body damage does not mean that a quale of pain entails knowledge. There can exist pain qualia without real body damage, for instance in the case of phantom limb pain (Flor 2002). Phenomenal information is not factive: there is nothing that prevents a signal from carrying false information. The point is that, for anything to be consciously known by a subject, it must necessarily be subjectively presented in some form. Qualia precisely refer to the forms of subjective experience. From this perspective, phenomenal information is our access to the world, it is a necessary condition for anything to be accessed. This is a crucial point for machine consciousness. What I am interested in when I wonder whether a machine can be conscious concerns its ability to access the world experientially. What I mean when I say that a robot has a conscious access to its body states is that it is informed of these states via a qualitative experience. If the machine accesses information with no qualia involved, it just means that the machine accesses it unconsciously. From this perspective, the well-known notion of "*access consciousness*" is ambiguous.

1.3. Access consciousness and cognitive qualia

Ned Block introduced a famous distinction between phenomenal consciousness (P-consciousness) that refers to what it's like to have an experience, and access consciousness (A-consciousness) that essentially corresponds to the informational and representational role of a conscious experience (Block 1995). A state is said to be A-conscious if one can use "*a representation of its content*" as "*a premise in reasoning*" and for "*the rational control of action*" and the "*rational control of speech*" (ibid.) For example, when Mary puts her hand on a hot plate and experiences pain for the first time, she is in an A-conscious state of pain in the sense that she can report feeling pain and control her behavior in order to remove her hand. P-consciousness refers to the qualitative aspect of experience. I use "phenomenal experience" and "qualitative experience" as synonyms. A quale of pain is a phenomenal experience of pain. Mary is in a P-conscious *state* of pain in the sense that she has a qualitative experience of pain.

How is P-consciousness supposedly distinct from A-consciousness? In order to highlight the difference, Block relies on the case of blindsight patients (Weiskrantz 1986). Also called "unconscious vision", blindsight is a neuropsychological phenomenon which results from lesions in the primary visual cortex. Blindsighters are able to discriminate certain visual stimuli, such as an "X" from an "O", while reporting that they don't really see them. They don't have an "X" that appears in their visual field, they just "guess" that it is an "X" rather than an "O". Real blindsight patients can guess what appears in front of them only when we ask them to do so, and if they have a small set of alternatives to choose from. Block imagines the case of a "*superblindsight*" patient who can guess whatever is in front of him while having no visual field:

> *The superblindsighter spontaneously says, "Now I know there is a horizontal line in my blind field even though I don't actually see it." Visual information from his blind field simply pops into his thoughts in the way that solutions to problems we've been worrying about pop into our thoughts, or in the way some people just know the time or which way is North without having any perceptual experience of it.* (Block 1995)

If this is a plausible scenario, Block argues that it would be an instance of A-consciousness without P-consciousness because the superblindsighter has access to visual information without having a visual experience. This case is interestingly similar to Mary's super-glass thought experiment, at least in one respect: both the superblindsighter and Mary have the ability to access information in an unusual way. Instead of visually accessing an "X" in front of him like other human beings, the superblindsighter has a "popping thought" that informs him of the presence of the "X".

The first important thing to point out is that the superblindsighter is in a P-conscious state when he has a popping thought. Indeed, if a state is said to be P-conscious when it is like something to have it, then it is hard to see why a "popping thought" does not involve P-consciousness. A good way to see that it is like something to have such a popping thought is that we, normal human beings, would actually be curious about what a superblindsighter experiences from a first-person perspective. Whatever it is like, the information that we access visually is accessed by the superblindsighter in another way. This implies, at least according to my definition, that superblindsight experiences are qualia. It may seem odd to talk about qualia in a case like this since the term is traditionally used in a more restrictive

sense. While it is widely accepted that there are perceptual or emotional qualia, the claim that propositional attitudes such as thoughts and beliefs can also involve qualia is more controversial. According to Michael Tye, the real bearers of the phenomenal character of propositional attitudes are sensations or mental images that accompany them:

> *On this view, in and of itself there is nothing it is like to remember that September 2 is the date on which I first fell in love. I might remember this fact (perhaps I read it in an old diary) but feel nothing at all. It is all ancient history to me now. No spark of feeling is produced.* (Tye 1995, p. 4)

There is no doubt that someone can be said to remember or believe something without P-consciousness being involved. However, this is not incompatible with the fact that there are thinking experiences or remembering experiences. Tye argues that if we "*take away the feelings and experiences*" that are associated with thinking or remembering, "*there is no phenomenal consciousness left*" (ibid.) Obviously, if we remove the experience there is no P-consciousness left, but it is not the point here. The superblindsighter undergoes thinking qualia in the sense that his subjective experience is modified in a certain way when looking at something, a modification that is characterized as a popping thought. The very idea of a *popping* thought entails that the thought *pops* in subjective experience, it manifests itself in subjective experience. An experienced popping thought is a quale because it is like something for the subject to have a thought. If I instantaneously become blind, deaf, insensitive to taste, smell, and to my own body, I will still have conscious thoughts and beliefs and they will be qualitative in the sense that it would be like something to have them.

In fact, according to the definition of part 1.1, every conscious experience is a quale since qualia refer to the way conscious experiences are subjectively lived. This is true in particular for conscious propositional attitudes such as conscious knowledge. Block compares superblindsight access to visual information to "*the way some people just know the time or which way is North without having any perceptual experience of it*". It could be objected that there is no phenomenology involved when one "just knows" something, not even an inner voice or a mental image. While I agree that an experience of "just knowing" is difficult to characterize precisely, it nonetheless has a subjective aspect that makes it a conscious experience. From a subjective point of view, there is an experiential change from a state of ignorance to a state of knowing, a feeling of certainty about the time it is or about which way

is the North. This kind of experiential change is particularly salient during an "*Eureka!*" moment, when the solution to a problem suddenly comes into mind. It would be arbitrary to consider that this experiential change is of a fundamentally different nature from that of pain. In both cases, there is "*what-it's-like-ness*" involved in a meaningful sense: a machine has neither visual experience nor popping thoughts. I will refer to such experiences of thoughts and beliefs as *cognitive qualia* (Shields 2011).

Block does not actually deny that superblindsighters are P-conscious since he writes: "*Of course, the superblindsighter has a thought that there is an X in his blind field that is both A-conscious and P-conscious"* (Block 1995). What he means is that "*the state of his perceptual system that gives rise to the thought*" is A-conscious without being P-conscious. The problem is that it is unclear why we should consider the cause of the thought as being conscious at all. Block argues that the content that is represented by the perceptual system is A-conscious. In other words, the information that is processed by the visual cortex of the superblindsighter is supposedly A-conscious without being P-conscious. I believe it is a confusing way of presenting the situation. As Bernard Kobes (1995) rightfully objects "*the availability is directly in virtue of the thought, and only indirectly in virtue of the underlying state of the perceptual system*". Visual information is consciously accessed through a popping thought and the only consciousness involved is due to this thought. If there is neither the manifestation of a visual field nor that of a popping thought in the experience of the superblindsighter, then it makes more sense to say that there is no consciousness of the visual information at all. In this view, A-consciousness is just not consciousness at all. This point is in fact made quite clear when Block writes this:

> *As an example of A-consciousness without P-consciousness, imagine a full-fledged phenomenal zombie, say, a robot computationally identical to a person, but one whose silicon brain does not support P-consciousness.* (Block 1995)

It is interesting that Block mentions robots as an example of "pure" A-consciousness. It is not difficult to imagine a robot reporting what is in front of him and using this information to control his behavior, all of this with no experience involved. In fact, we can imagine this because it is already achievable. Of course, today's robots are far from being "*computationally*" or behaviorally identical to humans. Nonetheless, we can imagine a future robot that is capable of doing everything a human does but has no conscious experiences at

all. Or can we? If we say that, we face a problem. If it is possible to imagine an unconscious robot that is indistinguishable from a conscious human being, why not also imagine that humans could possibly be "*zombies*"?

1.4. Zombies and illusionism

The "*Hard problem of consciousness*", as coined by David Chalmers (1995), consists in explaining how physical mechanisms give rise to subjective experience. Providing a complete explanation of consciousness constitutes a special challenge for science as it seems hard to understand how microphysical facts (about atoms, molecules and neurons) entail the existence of phenomenal consciousness. A popular argument against classical reductive explanations of consciousness appeals to the conceivability of zombies (Chalmers 1996, pp. 84-88). A premise of the zombie argument is that I can conceive a human being physically identical to myself who undergoes no qualia. This creature is called my zombie twin. Any physical or behavioral fact that is true about myself is also true about my zombie twin, he is molecule-by-molecule identical to myself and he says and does exactly the same things that I do. However, although he reacts the same way as I do when he burns his hand, my zombie twin does not feel pain because he never feels anything from a subjective point of view. His pain states are access conscious, thus he can report that he feels pain and behave accordingly but he does not really feel anything. There is nothing it is like to be my zombie twin. Since zombies are conceivable, so the argument goes, they are metaphysically possible. God could have created a zombie world devoid of qualia, that is yet physically identical to ours. As for Jackson's Knowledge argument, the zombie argument is designed to refute physicalism. Indeed, its conclusion implies that facts about the physical world cannot account for qualia. The zombie argument is controversial, as conceivability might not be a good guide to possibility (Yablo 1993). I will not discuss the validity and soundness of the zombie argument since my aim is not to refute physicalism for now. Nonetheless, I believe that thinking about zombies is a fruitful way to introduce the problem of qualia.

An interesting consequence of the premise that my zombie twin is behaviorally identical to myself is that this zombie will be as interested as I am about the Hard problem of consciousness. Indeed, doing philosophy and writing a master thesis about qualia are behaviors, and behaviors can be explained from brain mechanisms in cognitive science. They are "*easy problems*" according to Chalmers' classification because they only involve cerebral

information processing and behavior that do not pose methodological challenges for science. Now we have a problem: if a zombie can wonder about qualia without experiencing anything from a first-person perspective, this entails that my subjective experience is not the actual source of my questioning about consciousness. Hence, the implication of the metaphysical possibility of zombies is that real qualia play no role in one's beliefs about them. This is a point that is closely tied to what Chalmers calls the "*meta-problem of consciousness*" which is "*the problem of explaining why we think consciousness is hard to explain"* (Chalmers 2018). The meta-problem of consciousness is itself an easy problem as it merely consists in explaining oral *reports.* My zombie twin will *say* that the problem of consciousness is a hard problem. Like myself, he will manifest the intuition or the belief that there are qualia and that they are mysterious. Such oral reports are behaviors and thus they can in principle be explained by cognitive science. The reason why this point constitutes a problem is that it allows one to defend the following argument:

> *1. There is an explanation of our phenomenal intuitions that is independent of consciousness.*
> *2. If there is an explanation of our phenomenal intuitions that is independent of consciousness, and our phenomenal intuitions are correct, their correctness is a coincidence.*
> *3. The correctness of phenomenal intuitions is not a coincidence.*
> ---
> *4. Our phenomenal intuitions are not correct.* (ibid., p. 47)

The idea is that if beliefs about qualia can be explained without having to suppose the real existence of qualia, then there is no need to suppose that those beliefs are true. The conclusion of this argument is defended by illusionists such as Keith Frankish (2016). According to illusionism, we are zombies. When I conceive of my zombie twin, I am actually thinking about myself. Admittedly, I have the very strong intuition that there is something more about myself, but this intuition can be explained from brain mechanisms. Explaining this very strong intuition constitutes the "*illusion problem*" which is very hard but arguably easier than the Hard problem of consciousness (Kammerer 2016). When the intuition is explained, there is nothing more to account for, in exactly the same way as explaining one's intuition about ghosts does not require an additional explanation of what ghosts really are.

The situation is paradoxical: the zombie argument that is supposed to reveal the problem of qualia ends up showing that there is in fact no Hard problem of consciousness at all. So what went wrong? Can we seriously consider that qualia do not exist and that the problem that I am addressing in this work relies on a false intuition? I don't think so, at least not with the definition of qualia I am interested in. In my view, the main reason is this: a conscious experience of believing something is a cognitive quale. Let's suppose that I accept illusionism, I now believe that qualia do not exist. I can consciously experience this belief. Indeed, right now I am having this belief in mind, I can concentrate on it. It is like something to consciously believe that qualia do not exist. Non-human animals probably do not have such belief experiences and neither do robots. If conscious beliefs are qualia, then the content of the belief that "qualia do not exist" is self-contradictory. Therefore, illusionism is false. This argument is close to Descartes' "*I think, therefore I am*". I am saying that "*there is a conscious qualitative experience of belief, therefore qualia exist*". I can doubt about anything but not that I am consciously doubting. According to the way I defined them, there cannot be a conscious experience if there is no quale. This is because a quale is a qualitative experience by definition, it refers to the way an experience is. Now, an illusionist will say that I am just begging the question. Of course, if I am assuming that any conscious experience is a quale, illusionism is trivially false. But precisely, I think that any conscious experience is mysterious in the way an experience of pain is. What I want to explain is not just why it is like something to be in a particular state such as pain but why there is "what-it's-like-ness" at all. A conscious illusion is a qualitative experience. Machines do not have such experiential illusions. What I want to explain is how qualia are possible in this general sense, and there is no way qualia cannot exist in this sense, by definition. Humans realize qualia and thus zombies are not possible. Illusionists are in fact denying qualia in another sense, which corresponds to the traditional way of conceiving them: qualia as "phenomenal properties".

1.5. Qualia as real concrete processes

Before illusionists, Dennett also defended the idea that qualia do not exist. In an influential article called *Quining Qualia*, he proposed "*to destroy our faith in the pretheoretical or "intuitive" concept*" of qualia (Dennett 1988). His contention is that qualia are based on wrong intuitions that he tries to pump out by proposing a series of little scenarios designed to show that there are in fact no such things as these mysterious qualia. Dennett begins by

defining qualia as "*the way things seem to us*" which is not far from the way I defined them. Then, he precises the concept of qualia that he targets:

> *So, to summarize the tradition, qualia are supposed to be properties of a subject's mental states that are*
> *(1) ineffable*
> *(2) intrinsic*
> *(3) private*
> *(4) directly or immediately apprehensible in consciousness.* (Dennett 1988)

Before discussing those four properties of qualia, there is a very important point I need to discuss: the characterization of qualia as "*properties of a subject's mental states*" is incompatible with the way I defined them. Qualia cannot be properties. This claim may sound odd since most of the defenders of qualia usually characterize them as properties of subjects or of mental states. However, the traditional conception of qualia is not robust to Dennett's destruction. The root of the problem are to be found in a "*substance-oriented*" ontology that is the usual framework within which these debates take place, as Manzotti explains:

> *Like most of current philosophy of mind and neuroscience, the aforementioned definitions of qualia are based on a substance-oriented ontology. By substance-oriented, I mean any ontology that is based on individuals or substances like objects, people, or representations. Such ontological schema often use terms like "properties," "mental states," and "sensations," which refer to substance-like entities.* (Manzotti 2008)

The principal reason why qualia do not fit into such an "*ontological schema*" is that, if a quale is defined as the way one consciously accesses properties, qualia cannot be properties themselves. There is no possible conscious knowledge of any property of the world without a conscious experience of it. Qualia are our conscious access to the objective world, thus they certainly cannot have the same ontological status as that of the properties of the things that can be known.

Qualia are epistemologically prior to the object-property scheme of the world, as the latter is the result of high-level cognitive abilities acquired during childhood. This last claim should

not be viewed as a philosophical contention based on a controversial view of the world, but rather as a relatively original way to formulate an empirical fact known by psychologists, at least since the work of Piaget:

> *At first there is neither external nor internal world but a universe of "presentations" whose images are endowed with emotional, cenesthetic, and sensorimotor qualities as well as physical ones. This primitive universe constitutes thenceforth the child's self as well as the objective of his actions. Hence there are as yet neither substances nor individualized objects nor even displacements, since without objects changes of position cannot be distinguished from changes of state; there are only global events connected with movements of the body proper, hence with kinesthetic and postural impressions.* (Piaget 1937, p. 213)

While Piaget does not explicitly refer to the notion of qualia, what he describes as a qualitative universe of "presentations" corresponds to what I defined as a quale. It is important to realize that we, as conscious subjects, came to know about the objects and properties of the world thanks to the ability to have qualitative experiences that we were born with. As newborns, it felt like something to be in pain, and we learned the concept of pain thanks to the innate ability to feel it. More precisely, we learned to articulate the feeling of pain with other feelings such as visual stimuli:

> *Infants are born with the ready-made opportunity to link experiences from the various sense modalities, experiences that co-occur and tend to be qualitatively linked, corresponding to particular feeling tones and profiles.* (Rochat 2011)

In other words, our ability to represent substantial and stable objects with properties - what psychologists call "*object permanence*" (Baillargeon & al. 1985) - depends on a more fundamental ability to experience the world qualitatively. The problem of qualia is exactly that of making sense of this fundamental ability.

Rather than property-bearing substances, abilities are more appropriately interpreted as *processes*. By process, I simply mean a chain of events producing a result. When I am in pain, there is a real causal chain of events that produce a qualitative experience of pain. This qualitative experience of pain is what I call a quale of pain. The quale is not a property of

anything, it is the process itself. Now, this process can be said to have certain properties. A quale of pain is characterized by the property of being painful. Only a quale can have such a property. “Being painful” can be called a *phenomenal property*. Some processes of the world - that I call qualia - have phenomenal properties whereas others don’t. When I see a man hurting himself, I am seeing a process that has the phenomenal property of being painful[2]. By contrast when I see a robot hurting himself, I am seeing a process that has no phenomenal properties at all.

One important motivation of eliminativists and illusionists to deny the existence of phenomenal properties is that they find it suspicious that there could exist properties that only one subject can know. But this is not true: anybody can know as a fact that a process has phenomenal properties. This is what I discussed in part 1.2: even though Mary has never felt pain, she can know that pain experiences are painful. She can learn that in a book. Because she is an expert in the neurophysiology of pain, Mary even knows more about painful processes than I do. She can know that someone feels pain even if that person manifests no typical pain behavior, just by observing the brain activity of this person in a scan. What she doesn’t “know” is what pain feels like, but this is not the knowledge of a property of the world. When she feels pain for the first time, there is no new property of the world that she discovers. Instead, Mary acquires a new ability, the ability to feel pain. The reason why this ability is so special is because it enables her to consciously access real properties of the world. Such abilities are undoubtedly very hard to completely understand, but when they are defined like this there is no way one can deny their existence.

Another reason to be dubious about qualia and their phenomenal properties is the surprising fact that there could be properties that a subject can know without the possibility of error. For example, I cannot be wrong about the painfulness of my pain. How is it possible? This is a subtle point to which I will return. For now, I want to point out that it is not always the case that conscious subjects are certain about the phenomenal properties of their own qualia. Let’s take the example of a complex qualitative experience: the experience of love. There is something it is like to be in love. However, people can be in love without being sure about it. It is common to hear people say "*I feel something for this person but I don't know if it is love or not*". Being loveful is a phenomenal property of qualia but people that are in love may

[2] Or at least, I am partially perceiving the process.

have doubts about this property. La Rochefoucault wrote that some people would never have fallen in love if they had never heard of it[3]. What it suggests is that the way we characterize our experiences is partly arbitrary. We use words to talk about our experiences and we learned to use these words in social contexts. The example of love emphasizes that there is a difference between the real concrete qualitative experience and the properties we attribute to it[4]. It is the main reason why I don't want to define qualia as phenomenal properties. I think a lot of confusion comes from this. Saying that a quale is a real process happening in the world - the qualitative experience itself - makes it harder to deny. By contrast, phenomenal properties are abstract, they are labels that we use to characterize qualia. I really want to insist that qualia are not abstract at all according to my definition. In a sense, qualia are more concrete than anything else. From my subjective point of view, my feelings, thoughts, emotions and percepts constitute everything I concretely have. Thus it would be absurd to deny qualia in this sense. What is true however, is that we have no precise and rigorous way to talk about qualia. I agree with eliminativists like Dennett that the notion of phenomenal property is problematic in several respects. In the second part, it will become clearer why phenomenal properties are not "properties" in a straightforward sense. For now, my point is that the fact that we have no rigorous way to talk about qualia merely reflects our lack of a complete theory of consciousness.

Let's now get back to the four properties of qualia Dennett blames them for. Are qualia ineffable, intrinsic, private, and directly apprehensible in consciousness? As just said, the "direct apprehension" of qualia "in consciousness" is a delicate point to which I will get back to this in part 2.3. What about the three others?

First, qualia are indeed ineffable, they cannot be communicated by means of words. I cannot verbally express what it is like for me to drink a good wine. However, this does not mean that qualia cannot be reproduced:

> *If I want to communicate to Sara the taste of a certain wine, the only way is to make Sara reproduce the same process that took place when I tasted the wine. In other*

[3] "*Il y a des gens qui n'auraient jamais été amoureux s'ils n'avaient jamais entendu parler de l'amour*." (La Rochefoucault 1678, §136)

[4] What makes the matter complicated is that the way we think and talk about our experiences change their phenomenal character, and this is in fact how we could interpret La Rochefoucault's maxim. I will come back to this point in part 2.4.

> *words, I cannot explain to Sara how the wine tastes. However, I can convince Sara to drink it and thus go through the same process that constituted my phenomenal experience.* [...]
> *Qualia cannot be communicated using words, but they can be reproduced.* (Manzotti 2008)

Of course, when Sara drinks the wine, it is not the *same* process that happens. However, the two processes share an important property: they are gustatory processes. Being gustatory is a phenomenal property and we don't understand what it takes for a process to have such a property. Again, we don't know because we have no theory of qualia. Only qualia can have the property of being gustatory. Dennett writes that a "*wine-tasting machine*" could in principle be able to "*perform better than human wine tasters on all reasonable tests of accuracy and consistency the winemakers could devise*". And then he adds that:

> *surely no matter how "sensitive" and "discriminating" such a system becomes, it will never have, and enjoy, what we do when we taste a wine: the qualia of conscious experience! Whatever informational, dispositional, functional properties its internal states have, none of them will be special in the way qualia are. If you share that intuition, you believe that there are qualia in the sense I am targeting for demolition.* (Dennett 1988)

My claim is not that the machine will not reproduce gustatory qualia whatever its internal state. Rather, the point is that we have no idea what it would take for a machine to reproduce a gustatory quale, which is precisely the problem I am interested in.

Second, there can be several interpretations of what "intrinsicness" means when talking about processes. In Dennett' article, "intrinsicness" is used interchangeably with "non-relational". A process is typically relational in the sense that it involves different entities that entertain complex causal and structural relations. However, a quale of pain can be said to be "intrinsically" painful in the sense that it is painful in and of itself, independently of anything external to the process. In other words, a painful process is painful whether or not we interpret it as such. An experiential process such as pain is probably an extremely complex chain of causes and effects. What these causal chains are exactly is the core of the problem.

Finally, to say that qualia are "private" is just another way to say that they are subjective. When I feel pain, it seems that the feeling itself belongs to me, only *I* can feel it. Admittedly, the subjective character of qualia makes them very problematic for a rigorous scientific investigation. One reason why the problem of qualia is difficult is that we do not have clear methods and concepts to address subjective processes in a rigorous manner. However, one could argue that it is just an empirical fact that there are subjective processes and thus science should explain how subjectivity is possible based on objective observations. This is for example what John Searle writes on this:

> *Dennett has a definition of science which excludes the possibility that science might investigate subjectivity, and he thinks the third-person objectivity of science forces him to this definition. But that is a bad pun on "objectivity." The aim of science is to get a systematic account of how the world works. One part of the world consists of ontologically subjective phenomena. If we have a definition of science that forbids us from investigating that part of the world, it is the definition that has to be changed and not the world.* (Searle & al. 1997, p. 114)

So why not just accept that some processes are subjective and try to investigate them scientifically? In the next part, I will examine this path, which I will call *methodological dualism*. Then in part 1.7, I will show that this methodological stance cannot possibly lead to solving the problem of artificial qualia.

1.6. The correspondence problem and methodological dualism

When human qualia are postulated as an empirical fact, it becomes possible to investigate the relation between observable processes and human conscious experience. This is what cognitive scientists interested in consciousness actually do. For instance, the project of finding the "*neural correlates of consciousness*" (NCC) is an important research programme in the quest for a scientific theory of consciousness (Crick & Koch 1990, 1998 ; Koch 2004 ; Chalmers 2000). A central assumption of the NCC is that experiential processes systematically depend on brain activity. I will have the occasion to question this assumption later but let's suppose for now that it is true. The general method underlying the NCC is to observe brain activities of conscious human subjects while modifying the content of their experience by varying the stimuli to which they are exposed. That way, it is in principle

possible to *map* neural activity to conscious experience. For example, it could be observed that every time a human subject feels pain, the same neural network activates. The neural correlates of pain are defined as the minimal set of neural processes that are together sufficient for a human experience of pain to be realized. Of course, I am oversimplifying the problem here. In practice, consciousness scientists face a lot of methodological obstacles. Let's forget about methodological questions. For the sake of argument, let's suppose that we can discover all the precise neural correlates of pain. It would mean that we know exactly what brain processes correspond to painful processes in humans. This is what was supposed in Mary's superglass thought experiment: she became a super-expert in the neurophysiology of pain to the point that she knows exactly the type of pain that a subject is enduring just by looking at brain scans.

In a sense, Mary is able to explain qualia. If someone asks her why she started to feel pain after stopping her analgesic treatment, her answer would simply be that painkillers were changing the functioning of her brain. She feels pain because the neural correlates of pain are activated in her brain. It is an empirical fact that when a human brain is activated in a certain way, an experience of pain is realized. But then, one might be tempted to ask: "*why does this particular neural activity cause the feeling of pain?*". The intuition here is that the explanation is incomplete, even when all the NCC have been discovered. Indeed, a possible objection is that the NCC are "mere correlates" of conscious experience. The underlying idea is that the role of a theory of consciousness is not only to map observable processes to experiences, but to reveal the *nature of the relation* that consciousness entertains with brain activity. This objection has been anticipated long before the study of the NCC. In his book *The Logical structure of the world*, Carnap distinguished between what he called the "*correspondence problem*" and the "*psychophysical problem*":

> *The question is this: provided that to all or some types of psychological processes there corresponds a simultaneous process in the central nervous system, what connects the process with one another? Very little has been done toward a solution of the correlation problem of the psychophysical relation, but, even if this problem were solved (i.e., if we could infer the characteristics of a psychological process, and vice-versa), nothing would have been achieved to further the solution of the essence problem (i.e., the "psychophysical problem"). For this problem is not concerned with the correlation but with the essential relation; that is with that which "essentiality" or*

> *"fundamentality" leads from one process to the other or which brings forth both from a common root.* (Carnap 1928, §22)

According to Carnap, the psychophysical problem belongs to metaphysics. Metaphysicians have been arguing for centuries over this problem, some defending that physical states caused experience, others that the relation was a parallelism or an identity, but no progress has been made on this metaphysical question. Carnap suggests that if no consensus has never been reached, it may be because the question does not really make sense in the first place: the psychophysical problem is what he calls a "*pseudoproblem*". In his view, scientists should concentrate on the correspondence problem and forget the confusing idea that there is anything more to discover about consciousness than correspondence relations.

As a scientist, addressing the problem of correspondence implies committing to a form of methodological dualism. Methodological dualism is not a metaphysical position but rather a pragmatic stance, whose purpose is to avoid pseudoproblems. A relatively recent and influential example of methodological dualism is Francesco Varela's *neurophenomenology* that he presented as a "*methodological remedy for the Hard problem*" (Varela 1996). Neurophenomenology is explicitly inspired by the phenomenological tradition, as initiated in the beginning of the twentieth century by Edmund Husserl (1931). One central idea of phenomenologists is that it is not possible to study consciousness objectively like other phenomena of nature because consciousness is itself the source of all objective knowledge. From this perspective, it is impossible to reduce consciousness to processes happening in the natural world because what we call the natural world is in fact a theoretical construction that we abstracted from lived experience, what phenomenologists sometimes call the "*lifeworld*":

> *The lifeworld is, unsurprisingly, the world we live in. It is the world that we take for granted in daily life, it is the pre-theoretical world of experience, which we are all acquainted with, and which we typically do not question. Why does it need to be rehabilitated? Because the lifeworld has been forgotten and repressed by science, whose historical and systematic foundation it constitutes. Even the most exact and abstract scientific theories draw on the prescientific evidence of the lifeworld. In its search for objective knowledge, science has made a virtue of its ability to move beyond and surpass bodily, sensuous, and pragmatic experience, but has frequently*

> *overlooked to what extent it is enabled by those very same experiences.* (Zahavi 2018, chap. 4).

Based on this view, Varela claimed that consciousness scientists should not try to reduce experience to brain activity because it would be a mistake to believe that the latter are more fundamental than the former. The goal of neurophenomenology is to understand how "*phenomenological accounts of the structure of experience and their counterparts in cognitive science relate to each other through reciprocal constraints*" (Varela 1996). In other words, neurophenomenology aims to understand how experience is lived from the first-person perspective and then to find *correspondence relations* with brain processes as described by cognitive scientists. These relations will never be anything more than correspondence, and there is no need to ask further metaphysical questions.

As the neuroscientists Gerald Edelman and Giulio Tononi remark, "*unlike any other entity, [...] with consciousness* we are what we describe *scientifically*" (Edelman & Tononi 2013, chap. 2). The problem is not that we cannot explain consciousness because we are conscious ourselves. After all, we are living creatures but scientists can explain life. The problem is that, as just discussed, subjective experience is the process by which we access the world in the first place. Searle explains this quite clearly when he writes this:

> *We find it difficult to come to terms with subjectivity, not just because we have been brought up in an ideology that says that ultimately reality must be completely objective, but because our idea of an objectively observable reality presupposes the notion of observation that is itself ineliminably subjective, and that cannot itself be made the object of observation in a way that objectively existing objects and states of affairs in the world can. There is, in short, no way for us to picture subjectivity as part of our world view because, so to speak, the subjectivity in question is the picturing.* (Searle 1992, pp. 97-98)

The problem with methodological dualism is that there is no way it will lead to a solution to the problem of artificial qualia. The latter is in fact essentially a metaphysical question about the nature of experience, as I will argue in the next part.

1.7. Multiple realization and the problem of artificial qualia

Once the correspondence problem has been solved, why not just say that a qualitative experience is *identical* to its corresponding brain process? After all, if we have sufficient scientific proof that every time certain neural networks of the human brain activate then a pain experience happens, it is not unreasonable to infer that there exists a *psychophysical law* linking the two processes. This law may just be a law of nature. Of course, it is mysterious why there is such a law, but this is not a problem that is particular to neuroscience. It is a mystery why there are physical laws that govern the universe in general. Why is it true that energy and matter are nomologically related, as described by the equation $e = mc^2$? This is arguably a deep philosophical question. However, even without a clear answer to that, physicalism can be rationally defended. According to physicalism, there is nothing "*over and above*" the physical (Smart 1959). All the real processes of the world are physical in and of themselves, and the role of scientists is to understand how the physical world works. In the same way that scientists discovered that gravity is a deformation of spacetime or that life is a biochemical process, they can also discover that pain experiences are brain activities. We have no good reason to believe that humans are discontinuous with the rest of nature. We are made of atoms and cells and we know that we are the product of evolution by natural selection. Thus it would be strange to believe that the qualia that we realize are not physical processes. Of course, phenomenologists are right that experience is the source of knowledge and this undoubtedly introduces very important methodological obstacles for a complete science of consciousness. But why would this prevent us from defending a clear metaphysical picture of the world? There must be processes happening in the world that *are* qualia, and not just "correlated" to them. And it is hard to understand why these processes should be of a different nature from other non-experiential processes. As Smart writes:

> *You cannot correlate something with itself. You correlate footprints with burglars, but not Bill Sikes the burglar with Bill Sikes the burglar. So sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything should be explicable in terms of physics (together of course with descriptions of the ways in which the parts are put together - roughly, biology is to physics as radio-engineering is to electromagnetism) except the occurrence of sensations seems to me to be frankly unbelievable.* (Smart 1959)

The view that qualia are identical to brain processes is called the *identity theory* (Smart 2017). An identity theorist will say that a quale *is* a neurophysiological process. Neuroscientists may not have discovered everything about the brain yet, but in principle there is no reason to believe that they won't. If we had a complete theory of pain, we could say that such experience of pain is identical to such and such brain activation according to such psychophysical laws. In other words, brain processes can *realize* qualia, and this is just a metaphysical fact about our world.

The main issue with the identity theory comes from the distinction between a *token* quale and a *type* of quale (Wetzel 2018). As I insisted in part 1.5, what I call a quale is a particular qualitative experience happening in the world. I defined qualia this way so that it becomes very hard to deny their existence. When I have an experience of pain, there is a painful quale happening and I can hardly doubt that. Defined this way, a quale corresponds to what philosophers often call a *token* process. A token process concretely happens, it is a particular instance of pain. But what is pain in general? This is a difficult question. For example, it is very likely that cats feel pain. When we see an injured cat exhibiting typical pain behavior, most of us have the intuition that the cat is really feeling pain. In other words, a real token quale is realized when a cat is injured. Now, what do I mean when I say that the cat realizes a *pain* quale? It seems that what I am trying to say is that there is an important sense in which the experience of the cat is similar to my experience. Let's call that *phenomenal similarity* (Shoemaker 1975). The gustatory qualia that are realized when I drink wine are phenomenally similar to the ones other people realize when they drink wine. In the same way, saying that a cat feels pain means that the quale the cat is realizing is phenomenally similar to my pain experiences. And this is how the so-called "phenomenal properties" are usually conceived: they are constituted by what phenomenally similar qualia have in common. Once we say that qualia have phenomenal properties, we can abstractly construct the class of all qualia that share a particular property and call this a *type* of quale. For example, all the processes of the world that are similar to what I call my pain qualia constitute one type of quale, namely pain qualia. They share the phenomenal property of being painful and it is this property that philosophers traditionally call a quale.

Accepting that there are types of qualia introduces a major problem known as multiple realization (Putnam 1967 ; Bickle 2020). According to the multiple realization thesis, the

same type of quale can be realized by different physical processes. This is what I just discussed: the pain quale realized by a cat is realized by a physical process that is different from the processes that realize my pain quale. So now, we must find psychophysical laws linking types of physical processes to types of qualia. And we must say that certain types of physical processes are identical to types of qualia. This position is called *type identity theory*, or "type-type identity theory" (Jackson & al. 1982) and it is close to *functionalism,* which I will present in the third part of this work. The type identity theory is problematic because it is very hard to see how we could possibly find psychophysical laws linking physical properties to phenomenal properties.

To understand why it is hard to find such psychophysical laws, let's suppose again that we have solved the correspondence problem: we know exactly which *human* brain processes correspond to pain. For example, neuroscientists could discover that the activation of a neural network happening in the neocortex is systematically correlated with human pain. A type identity theorist could thus infer the following psychophysical law: if such neural network of the neocortex activates, then a pain quale is realized. Of course, I am here oversimplifying what a psychophysical law could be. The essential point is that we could have a type of physical process that is nomologically linked to a type of quale. Such a law would be very useful because it could allow scientists to infer truths about the experiences of non-human animals. Indeed, if cats also have a neocortex that activates in a similar way when they exhibit a typical pain behavior, it would be rational to infer that they really feel pain. Moreover, this law could also lead scientists to clearly make a distinction between creatures that feel pain and those who don't. For example, do fish feel pain? It turns out that, unlike cats, fish do not have a neocortex. Thus the type identity theorist could infer that the body of a fish does not meet the sufficient conditions to realize pain qualia. This is what Dinets (2016) and Michel (2019) call the "*no cortex no cry argument*":

> *No cortex, no cry argument:*
> *(1) If x feels pain, then x has a neocortex*
> *(2) Fish do not have a neocortex*
> *(3) Therefore, fish do not feel pain*
> (Michel 2019)

The issue is that there is no way to be sure about *(1)*. "*If x feels pain, then x has a neocortex*" is a negative psychophysical law that we have no way to test. Once we accept multiple realization, why not suppose that very different physical structures that do not involve a neocortex can also realize pain? If very intelligent aliens come to visit us, learn our language, and report having very unpleasant pain experiences when they get hurt, we will believe that they realize pain qualia. And we will believe it even if they have no neocortex. So why not suppose that fish realize pain with a very different brain activity than ours? Some fish do manifest pain-like behavior, but unlike humans and intelligent aliens they are unable to report their experience. The problem is that we have no general criteria to decide if a creature feels pain when it is unable to report its own experience. And we cannot simply say that a sufficient condition to realize pain is to exhibit a pain-like behavior because a simple artifact can do that:

> *For example, one could build a simple robot programmed to do just three things: when exposed to a noxious stimulus, it moans and shouts, protects the area of its body exposed to the noxious stimulus, and runs away. Despite the fact that the robot exhibits pain-like behaviors, attributing pain to this robot would not be the most rational thing to do.* (Michel 2019)

Now we start to see how the problem of artificial qualia is intimately linked to the multiple realization problem. To build a conscious machine, we must find all the necessary and sufficient conditions that a physical process must meet to instantiate a phenomenal property such as "being painful". But there is no way scientists will find these conditions because they must have detection procedures to find correspondences between experience and qualia. For humans, it is in principle possible to solve the correspondence problem because they can report their own qualitative experience. Because of the possibility of multiple realization, it is impossible for scientists to infer general psychophysical laws that would tell us without doubt whether or not a non-human animal or a machine is conscious. This is because scientists are stuck in a vicious circle: they need to find criteria to distinguish an experiential process from a non-experiential process but they need these criteria in the first place to detect consciousness. Matthias Michel summarizes this circularity quite clearly:

(1) To find a criterion that demarcates between plausible and implausible cases of multiple realizations of consciousness, one must provide a necessary condition for consciousness.

(2) To know whether a condition C is necessary for consciousness, one must test hypotheses of the form: for all entities, if E is conscious, E satisfies C.

(3) In order to test these hypotheses, one both needs to know whether the tested entity E satisfies C, and whether E is conscious or not.

(4) However, we do not know if E is conscious or not, since it is precisely what we try to assess.

(5) Therefore, we cannot find a criterion that demarcates between plausible and implausible cases of multiple realizations of consciousness.

(Michel 2019)

What conclusions should we draw from this? First, Carnap was wrong when he claimed that the psychophysical problem is a pseudoproblem. Whether a non-human creature like a fish or a robot is capable of feeling pain is a substantial question. It is interesting to note that Carnap considered the case of artificial experience. He claimed that the following question is meaningless when expressed as such: "*if a robot is exhibiting all the behavior appropriate to tooth-ache, is there a pain connected with that behavior or not ?*" (Carnap 1936). If this question is meaningless, then we must figure out how to give it a precise meaning. The second conclusion that we can draw is that science alone cannot solve the problem of artificial qualia. Thus Carnap was right when saying that the psychophysical problem concerns philosophers. It seems that the root of the problem comes from these notions of "phenomenal property" and "types of experience". What we mean when we ask whether or not a non-human creature really feels pain is not clear at all. In the second part of this work, I will try to understand why it is so difficult to talk rigorously about the phenomenal content of a conscious experience. I will start with the philosophical problem of perception. As I explained, qualia are our access to the external world: according to the way I defined them, there is no conscious perception if there is no qualia. However, most cognitive scientists ignore this concept of qualia. The most common way to understand perception is in terms of mental representations. I will discuss representationalism and try to make sense of qualia in a representationalist picture of perception. Then I will present adverbialism and show why it is difficult to abstract "phenomenal properties" from experience. In the end of the second part, I will examine the possibility of rigorously modelling the phenomenal contents of experience.

2. An analysis of qualia

2.1. The representationalist view of perception

A straightforward way to introduce the problem of perception is to ask the following question: when I consciously perceive a real object, do I perceive it as it "really is" or do I perceive something mediating between me and the object as it is "in itself"? To address this question, let's outline a simple scheme of the act of looking at a red apple. First there is the red apple itself, a physical object made of particles, as described by physics. Then there is myself, the perceiver, equipped with a visual system essentially composed of two eyes connected to a brain. In order for me to see the apple, there must also be light waves reflected by the surface of the apple reaching my retinas. After being captured by my retinas, the light waves are converted into neural signals. At this point, very complicated processes happen in my brain. Regardless of the details, what we know for sure is that it is in large part because of these brain processes that conscious seeing of the apple is possible. Now, the problem is the following: if what I see is the result of my brain activity, can I say that it is really the apple that I am seeing? This question is of course formulated in a deliberately ambiguous way, yet I believe it adequately points to the core of the problem of perception. The underlying intuition is that if what I see is the result of something that is happening inside my head, then it follows that the object of perception is some sort of internal construction. Indeed, the same brain processes could happen without there being any real apple in front of me, in which case I would be in a state of hallucination. In the case of hallucination, as well as in the case of veridical perception, there is a visual experience of an apple. Thus it seems that the object of the experience is not a real apple after all. This is why I may be inclined to believe that I never see anything *directly*.

The complete formal argument leading to the thesis of *indirect perception* goes as follows:

> *1. The hallucination and the veridical experience can be type identical. This is why the possibility of hallucination is so distressing. The perceiver herself could never tell a difference just from the character of the experience whether it was veridical or hallucinatory.*

2. Because they are identical you have to give the same analysis of each.
3. But in the hallucinatory case you do not see a material object.
4. But you do see something. There is no question that this is an instance of visual perception. Hallucinatory or not, there is a seeing of something.
5. But the something is not a material object. Give it a name, call it a "sense datum".
6. But by 2. you have to give the same analysis of each, so if you don't see a material object in the hallucinatory case, you do not see one in the veridical case. In both cases you see sense data. Indeed, all we ever perceive are our sense data. (Searle 2018)

Searle calls this the "*Bad Argument*". According to him, most of the great Western philosophers of the modern era ("*Descartes, Leibniz, Spinoza, Locke, Berkeley, Hume, and Kant*") defended the thesis of indirect perception, and they typically based their view on a variant of this line of reasoning. It is a bad argument because it relies on an ambiguous use of the verb "to see". In the ordinary sense of the word, I do not really "see" anything in the case of a hallucination. Rather, I falsely *believe* that I see something. That said, it remains true that a hallucination is a visual experience. Thus the problem is now to characterize the content of a hallucinatory experience. Searle mentions "*sense-data*", a concept that was notably developed in the beginning of the twentieth century by Bertrand Russell :

Let us give the name of 'sense-data' to the things that are immediately known in sensation: such things as colours, sounds, smells, hardnesses, roughnesses, and so on. We shall give the name 'sensation' to the experience of being immediately aware of these things. Thus, whenever we see a colour, we have a sensation of the colour, but the colour itself is a sense-datum, not a sensation. The colour is that of which we are immediately aware, and the awareness itself is the sensation. (Russell 1912, p. 4)

This is an explicit and clear version of indirect perception. In veridical and non-veridical cases of perception of a red apple, the redness that I am aware of is the same thing: a red sense-datum. Sense-data provide a simple answer to the question of the content of hallucinatory experience but it introduces a problem that Russell formulates right after:

> *It is plain that if we are to know anything about the table, it must be by means of the sense-data—brown colour, oblong shape, smoothness, etc.—which we associate with the table; but, for the reasons which have been given, we cannot say that the table is the sense-data, or even that the sense-data are directly properties of the table. Thus a problem arises as to the relation of the sense-data to the real table, supposing there is such a thing.* (ibid.)

Any view that postulates experiential intermediaries faces the problem of the relation between what belongs to the "internal" realm of experience and what belongs to the "external" world. This problem is particularly salient with the sense-data theory since it suggests the idea of a "veil of perception" between the mind and the world. Because of the huge epistemological and ontological problems they pose, sense-data became unpopular in contemporary philosophy of mind.

Since the rise of cognitive science in the second half of the twentieth century, many theories of conscious perception have been developed in terms of *representations* (Lycan 2019). The essential claim of representationalism is that perceiving the redness of the apple amounts to being in a certain *state* that represents the apple as being red. Such a state is called a *representational* state. Fred Dretske introduces representationalism as follows:

> *If, in accordance with the Representational Thesis, we think of all mental facts as representational facts, the quality of experience, how things seem to us at the sensory level, is constituted by the properties things are represented as having. My experience of an object is the totality of ways that object appears to me, and the way an object appears to me is the way my senses represent it.* (Dretske 1997, p.1)

What Dretske means is that the redness that I am aware of in a visual experience is the property that the visual experience attributes to the apple. Visual experiences are all about attributing properties to external objects. In this representationalist picture of perception, there are no explicit intermediaries between the perceiver and the real object. My visual system represents the apple as being red but there is not literally "something" red in my experience. This is because a thing that represents does not need to have the properties of the thing represented. For instance, a stationary radar can represent a car as going at 100 km/h without having the property of moving itself:

> *According to a representational theory, experiences (of movement, say) are like that. The representational vehicle, the thing in your head, doesn't (or needn't) have the properties (movement) it represents the world as having.* (Dretske 2003)

As Dretske explains, a perceptual experience involves a complex physical state of my brain, called the *vehicule* of the representation. This brain state has the ability to represent a property of the world and this represented property constitutes the *cognitive content* of the representation. For example, when I see a red apple, there is a brain state in my visual cortex that represents the color of the apple. It is in virtue of this brain state that I can know that the apple is red. The representationalist view is advantageous because it keeps intact the idea that perceiving essentially consists in detecting properties of the world while leaving open the possibility of perceptual error. Indeed, representationalism can easily account for misrepresentation. When I am in a hallucinatory state, I am simply misrepresenting my external environment as being in a certain way, maybe because my visual system is malfunctioning. Hence, representationalism appears to be a good framework to understand perceptual experiences. There remains of course a lot of details to explain and it is the role of cognitive scientists to develop complete models of perception, detailing precisely how exactly the brain represents properties of the world.

2.2. What is a mental representation?

The issue with representationalism as just presented is that it is very hard to see how it is going to help understand how *conscious* perception is possible. It is already possible to build machines with the ability to form complex visual representations of their environment. Self-driving cars are good examples of artifacts that can perceive the world and react accordingly thanks to internal representations that are physically realized by electrical circuits. However, almost nobody would claim that self-driving cars consciously experience the world. Representationalists would either argue that their perception is unconscious or that it is not even a case of perception, depending on the way they define the term. So what is needed to consciously perceive the world? There are essentially two types of representationalist answers to this question.

The first type of answer relies on *higher-order theories* (HOT) of consciousness:

> *Higher-order theories of consciousness argue that conscious awareness crucially depends on higher-order mental representations that represent oneself as being in particular mental states.* (Lau & Rosenthal 2011)

According to HOT, it is only when a subject is explicitly aware of a mental representation that the latter becomes conscious. The example of the absent-minded driver illustrates this point (Armstrong 1968). Sometimes, when we drive on an empty highway, we get lost in our thoughts and our attention is no longer explicitly directed to the visual information coming from the road. In this case, HOT theorists will say that the driver perceives the road unconsciously. Then, to explain how to go from an unconscious representation to a conscious one, they appeal to more complex representational states (thoughts) whose contents are themselves representations. It is in this sense that they talk about "higher-order" states: it is not enough to have a first-order representation in order to perceive consciously, there must be at least a second-order representational state that is directed to the first-order one.

There exists a second type of representationalist views that aim to account for experience in a first-order sense. A famous objection to higher-order theories is that they seem to entail that most animals are not conscious since they may not have the complex cognitive abilities required by HOT (Gennaro 2004). Some of us have the intuition that it is like something to perceive the world as a cat even if the cat has no thoughts at all and is not explicitly aware of its own percepts. Tye (2002) proposes a representationalist theory of conscious perception that, according to him, could apply to simple minds such as that of honey bees and fish. His central claim is that the phenomenal content of experience is reducible to certain sort of representational content:

> *The best hypothesis, I suggest, is that visual phenomenal character is representational content of a certain sort—content into which certain external qualities enter. This explains why visual phenomenal character is not a quality of an experience to which we have direct access (representational content is not a quality of the thing that has representational content) and why visual phenomenal character necessarily changes with a change in the qualities of which one is directly aware (changing the qualities changes the content).* (Tye 2002)

According to Tye, the qualitative aspect of experience is inherited from real external qualities. These external qualities, such as colors, can enter into the content of some representational states. But how is that possible? The question is precisely that of determining what enables qualities to enter "into" the cognitive content of representations.

An important common point between the first-order and higher-order theorists is that both talk about representational states that they already suppose are *mental*. What makes a representational state a *mental* state? One very simple way to think of a mental representation is in terms of information. A mental representation is something that a cognitive system possesses, and in virtue of which the system can use the content of the representation to do things. However, this definition is not restrictive enough. A simple system such as a thermostat could be said to have a mental representation with this definition. Indeed, a thermostat is designed to be in a state that represents a temperature and use the cognitive content of the representational state to do something (cooling or heating a room). Almost nobody would claim that a thermostat has a *mental* representation of the temperature[5]. We generally do not attribute mental representations to artifacts because we know they have no mental lives. But what does it mean to have a mental life?

It seems that a necessary condition for a creature to have a mental life is that it has the capacity to experience the world. A cat has mental representations because it is like something to be a cat. If, as Descartes believed, non-human animals are mere automata, it would be arbitrary to say that cats have mental lives whereas thermostats don't. Indeed, any definition of a mental representation that is only based on information processing will have trouble distinguishing between mental and non-mental informational processes. Now, that does not mean that the content of a representation needs to be consciously accessed to be called mental. At least since Freud, we know that most of our mental lives take place unconsciously. It is undeniable that there are unconscious mental representations such as unconscious desires and unconscious beliefs. Moreover, there exist cases of unconscious perceptual representations such as subliminal percepts (Merike 2000). So what distinguishes an unconscious mental representational state from a representational state that is not mental at all? One way to make sense of this distinction is to say that a representational state is mental

[5] Although Chalmers defends a *panpsychist* view according to which a thermostat could have mental representations (Chalmers 1996, chap. 8.4). I will get back to panpsychism in part 3.5.

if its cognitive content has the *disposition* be experienced consciously by the system that is in this state. This is what is proposed by Searle:

> *All my mental life is lodged in the brain. But what in my brain is my "mental life"? Just two things: conscious states and those neurophysiological states and processes that—given the right circumstances—are capable of generating conscious states.* (Searle 1992, chap. 7)

We consider that beliefs and desires are mental representations even when they are unconscious because we know that their content can in principle be consciously experienced by the one that has it. As Searle (ibid.) explains, Freud himself believed that all mental states are "*unconscious in themselves*" (Freud 1915). Freud thought that unconscious mental representations were like objects that we can "perceive", to bring them to consciousness. In this respect, his interpretation of consciousness is close to that of contemporary HOT. The crucial point is that, on this view, conscious and unconscious mental representations presupposes consciousness. Artifacts have no unconscious mental representations because they have no mental representations at all.

Of course, it is possible to deny this and define mental representations in another way. It is also possible to accept that robots or simpler artifacts such as thermostats have mental representations. However, I believe it would be a very counter-intuitive use of the term. When I talk about my mental representations, I typically think of the representations that can be presented in my subjective experience. This is what I mean when I say that I want to build a conscious robot: I want it to be able to represent properties and to make these properties *present* to its own subjective experience. Thus, the problem of artificial experience is not that of finding how to go from unconscious to conscious mental representations. Instead, it is a deeper question that concerns content that can be *present* in an experience. In the next part, I will argue that the content of a conscious experience consists in *modes of presentation* of represented properties.

2.3. Perceptual modes of presentation and phenomenal concepts

The notion of mode of presentation was introduced by Frege (1948) in his theory of concepts. Frege makes a distinction between the sense and the reference of a linguistic expression: "*the*

morning star" and "*the evening star*" both have the same *referent* or *extension* (the planet Venus), to which they refer in different ways. "*The morning star*" and "*the evening star*" are two modes of presentation of the same referent. Chalmers (2004) characterizes a mode of presentation as a condition on extension. The concept "*morning star*" is associated with a condition like: "*the object usually visible at a certain point in the morning sky*". In our world, Venus satisfies this condition, so Venus is the extension of this concept and it is also the extension of the concept "*evening star*", even though it has a different condition on the extension. Chalmers proposes to extend this approach to the content of perceptual experiences: "*Perceptual experiences attribute properties to objects: e.g., my visual experience might attribute greenness to a ball*" (ibid.) This is the main idea of the representationalist view of perceptual experience: it represents objects as having certain properties. Then, Chalmers explains the notion of *perceptual* mode of presentation:

> *At a very rough first approximation, one might say that for a property (say, greenness) to be attributed by the experience, it must be the property that has usually caused that sort of color experience in normal conditions in the past. So the mode of presentation of the property will be something like:* the property that usually causes phenomenally green experiences in normal conditions. (Chalmers 2004, p. 24)

The idea is that, just like a linguistic expression refers to its extension in a certain way, a perceptual experience also represents a property in a certain way. The perceptual mode of presentation of the property "being green" is the phenomenal character of normal experiences of green things. Chalmers calls this the "*Fregean content*" of experience, as opposed to the "*Russellian content*" that representationalist theories such as Dretske's and Tye's imply. Brad Thompson explains this difference quite clearly when explaining that Russellian theories of phenomenal content are those that accept the following "*Russellian thesis*":

> *For any experience (that has phenomenal content) with phenomenal character r, there is some property p_r such that, necessarily, if an experience has phenomenal character r then it attributes p_r.* (Thompson 2009)

This Russellian thesis means that the phenomenal content of experience concerns uniquely *what* properties objects are represented as having. For example, when I see a red apple, redness appears in my visual field. According to Russellian representationalism, the redness

is the property my visual experience represents the apple as having and everything about this color experience concerns only *what* is represented. Alternatively, "*a content is Fregean if it consists of modes of presentation of objects and properties rather than the objects and properties themselves*" (ibid.) According to Fregean representationalism, the phenomenal content experience concerns *how* the world is represented rather than *what* is represented. When I see a red apple, my experience represents the apple as being red, and it does it in a certain way. This way that the redness of the apple is represented is a perceptual mode of presentation of the represented redness and it corresponds to phenomenal redness. Similarly, when Mary feels pain for the first time, she represents her body as being damaged in a new way. The painful character of her experience is a new mode of presentation of her body damage.

Let's take the example of the property "being a sphere". When I look at a ball, the sphericity of the ball is represented by my visual system. Because I have a *conscious* visual experience of the ball, I have a qualitative experience with a certain visual phenomenal character *V*. Now, if I touch the same ball with my eyes closed, I also have a conscious experience of the sphericity of the ball with another phenomenal character *T*. According to Fregean representationalism, the same property is represented in two ways, and these two ways correspond to two perceptual modes of presentations. Can we also make sense of this difference with Russellian representationalism, only in terms of *what* is represented? According to Dretske, we can:

> *in representing F-ness in mode V (vision) and T (touch), the phenomenal difference in our awareness of F-ness might be explained as the difference in representing F and (some aspect of) V in the first case and F and (some aspect of) T in the other. In representing a property, there is—or there may always be— a representation of the channel over which information about that property is received.* (Dretske 2003)

What Dretske means is that the way the information is represented is itself represented in perceptual experience. At first glance, this might seem like a valid objection to Fregean representationalism. After all, I know that I accessed a visual representation *visually*, just by having a visual experience. Thus it may be that the visualness of a visual experience is itself represented in the visual experience. The problem is that it is very difficult to identify the phenomenal character that corresponds to the visualness of a visual experience. This is

because the visualness of a visual experience is not something that is represented. This would be equivalent to saying that a picture represents its "pictureness". A picture does not *re*present anything about itself. Instead, the picture represents something else by simply being present. In the same way, the visualness of a visual experience is what is presented in visual experience. It is impossible to conceptually grasp what this purely "*presentational content*" of experience is because it is "*cognitively unavailable*" and "*nonconceptual*" content (Metzinger 2004, chap. 2.4.4). The ineffability of its pure "*suchness*" makes it impossible to characterize it precisely (ibid., p. 94). It is this presentational content that is sometimes characterized as being composed of individual phenomenal "objects" such as sense-data. For example, if there is a very particular shade of turquoise that stays the same in my visual field, it is tempting to say that *this* shade that I am aware of is a simple atomic sense-datum of turquoise. As Thomas Metzinger explains, this would be a mistake:

> *Even if simple presentational content, for example, a current conscious experience of turquoise$_{37}$, stays invariant during a certain period of time, this does not permit the introduction of phenomenal atoms or individuals. Rather, the challenge is to understand how a complex, dynamic process can have invariant features that will, by phenomenal necessity, appear as elementary, first-order properties of the world to the system undergoing this process.* (Metzinger 2004, p. 94)

The property "*being turquoise$_{37}$*" is a property that something is represented as having by my visual experience. Thus it would be a mistake to say that it is my visual experience itself that has this property. I will get back to this just after. For now, the important point is that, although it is very difficult to talk about it, there is a presentational content of visual experience that is very difficult to characterize.

One way to intuitively grasp this presentational content is to think of an *inverted spectrum* scenario (Locke 1689 ; Shoemaker 1982). I could wake up tomorrow and see all the colors inverted: apples look violet, the grass looks blue and the sky looks green. Maybe it is because of a brain disorder, or maybe I just put special lenses. At first, I would have a hard time talking about the colors of objects. In fact, I will be in a constant state of misperception. I will say that the grass looks blue because I always used the word "blue" when having an experience with this particular phenomenal character. Over time, I will eventually relearn to use the right words to refer to the colors of objects. I will understand that when I have a

phenomenal experience of green, I must say blue and vice versa. What changed here are the modes of presentation of the represented colors. Let's imagine now that after I have adapted to my new situation, I am able to represent the colors of objects correctly. I am able to have veridical visual experiences again. Although my experience still has a blue phenomenal character when I look at the grass, I represent it as being green because I know that experiences that had this blue phenomenal character in the near past were caused by green objects. One may object that my new situation will not result in exactly the same representational states as before. Maybe the human visual system is able to discriminate more shades of green than shades or blue for example. If this is the case, the inversion will introduce perceptual mistakes that I was not making before: now when I see a turquoise object I will say that it is green whereas I was saying that it is blue before my inversion. But would it really be a mistake? Such disagreements about the colors of objects are common between people. Labels that we use to talk about colors are partly arbitrary and it can vary from one culture to another. What matters is that, in most cases, people agree on the way to label colors. Thus, not only is it conceivable that my spectrum could be inverted, but it may be that different people actually have very different phenomenal experiences when looking at the same object. In fact, I could have been born with this inversion and I would never have noticed. Russellian representationalism has difficulties to account for these types of conceivable scenarios because it denies the existence of perceptual modes of presentation. If everything in visual experience is about how objects *are*, it is difficult to see how different phenomenal experiences could represent the same property.

The problem now is that saying that a visual experience contains modes of presentation of properties rather than properties themselves seems to get us back to indirect perception. Aren't we saying the same thing as the sense-data theorists, namely that we see intermediary perceptual objects instead of the objects themselves? At first glance, Fregean representationalism is incompatible with the "transparency of experience":

> *Focus your attention on a square that has been painted blue. Intuitively, you are directly aware of blueness and squareness as out there in the world away from you, as features of an external surface. Now shift your gaze inward and try to become aware of your experience itself, inside you, apart from its objects. Try to focus your attention on some intrinsic feature of the experience that distinguishes it from other experiences, something other than what it is an experience of. The task seems*

> *impossible: one's awareness seems always to slip through the experience to blueness and squareness, as instantiated together in an external object. In turning one's mind inward to attend to the experience, one seems to end up concentrating on what is outside again, on external features or properties.* (Tye 1995, p. 30)

For sense-data theorists, the transparency of experience is problematic because if visual experience was really presenting intermediary objects of perception, it is hard to understand why we always see external objects having the perceived properties. However, this is not a problem if visual experience contains modes of presentation. This is because we see *through* modes of presentation, as Chalmers explains:

> *When one introspects the content of a belief such as* Hesperus is bright*, one does so by thinking about Hesperus; one looks right through the mode of presentation. But nevertheless the mode of presentation exists, and one can become introspectively aware of it.* (Chalmers 2004, p. 28)

In the same way, we look right through perceptual modes of presentation. When I look at the red apple, it is very hard to "unsee" that the redness I am conscious of belongs to the apple. However, it is not impossible to become aware of the perceptual mode of presentation itself. In introspection I can focus my attention on the way my experience presents redness to me, in the same way that I can focus my attention on the word "cat" itself instead of thinking about a real cat. And when I become aware of the mode of presentation itself rather than the represented property, I refer to it by using a *phenomenal concept*.

A phenomenal concept is a very particular kind of mental representation. The peculiarity of a phenomenal concept is that it refers to a conscious experience. Not just any conscious experience, but a conscious experience lived by the subject using the concept. For example, let's suppose that I feel pain right now. In introspection, I can recognize the feeling of pain and refer to this feeling by using the concept of pain. This concept of pain is a phenomenal concept: it refers to a conscious experience I am having. Brian Loar characterizes phenomenal concepts as "*recognitional concepts*":

> *They* [recognitional concepts] *have the form 'x is one of that kind'; they are type-demonstratives. These type-demonstratives are grounded in dispositions to*

> *classify, by way of perceptual discriminations, certain objects, events, situations.* (Loar 1997)

For example, I can have a recognitional concept of pine trees. Every time I see a pine tree, I recognize that it belongs to the kind "pine tree" in virtue of a recognitional concept that is activated in my mind. In the same way, every time I feel pain, I recognize in introspection that it is "pain" that I am feeling and thus I also have a recognitional concept of pain. However, there is an important difference between the recognitional concept of pain and that of pine trees. When I recognize a pine tree, I have a visual experience with a certain phenomenal character that I usually have when I see a pine tree: the property "being a pine tree" is represented in my visual experience through a certain perceptual mode of presentation. By contrast, when I recognize pain in introspection, my recognitional concept of pain immediately grasps its referent. As Loar puts it, an experience "*serves as its own mode of presentation*" (Loar 2004, p. 229). In introspection, the way the concept of pain refers to pain is to present pain itself. Thus a phenomenal concept is in "direct contact" with its referent. Or to use a metaphor of Katalin Balog (2009): "*a token of the reference provides the ink in which the token concept is written*". This is why the phenomenal character of experience is "*immediately apprehensible in consciousness*", as Dennett (1988) puts it. However, for a physicalist, all of this is very mysterious. Loar suggests that the referents of phenomenal concepts are neural states. In other words, what we recognize as "pain" in introspection *is* a property of the brain. The problem is that it is very hard to understand how this painfulness that is directly apprehensible in introspection could possibly be identical to a neural property. The idea that neural states can present themselves painfully in introspection in such an immediate way is difficulty intelligible[6].

It is very difficult to analyze the phenomenal content of experience because a phenomenal character such as phenomenal redness cannot be straightforwardly interpreted as a "phenomenal property". As explained in part 1.7, the intuition behind the notion of phenomenal property is that there are *phenomenal similarities* between experiences. My experience of red right now is similar to the experience of red I had yesterday, thus it is tempting to say that the two experiences share the phenomenal property "being red". The problem is that it is this phenomenal similarity that allows us to represent something as being

[6] Kammerer (2019, chap. 7) offers a detailed physicalist critique of Loar's theory, as well as similar theories of phenomenal concepts such as Balog's (2012).

red. According to Fregean representationalism, the visual experience of a red apple contains a mode of presentation that is defined with respect to past experiences. The phenomenal character caused by red objects in the past is similar to the one my experiences has right now, thus I represent the apple as being red. Hence, "being red" is always a property that is attributed to something that the experience represents, it is not a property of the experience itself. The phenomenal character of an experience is not a property of anything. In the next part, I will argue that the phenomenal character of experience is more appropriately interpreted as an *adverbial* modification of a holistic qualitative experience.

2.4. Adverbialism and holistic qualia

The main idea underlying the adverbial analysis of perception, first introduced by Ducasse (1942) and Chisholm (1957), is summarized by Jackson as follows :

> *The basic idea behind this analysis is to utilise the fact that, on standard views, appearances, after-images, sense-data, and so on, cannot exist when not sensed by some person (sentient creature), in order to reconstrue statements which purport to be about appearances, after-images and so on, as being about the way or mode in which some person is sensing. Hence a statement of the form 'x presents a red appearance to S', becomes 'S senses red-ly with respect to x', and 'S is having a square sense-impression' becomes 'S is sensing square-ly'.* (Jackson 1975)

The advantage of adverbialism is that it avoids reifying appearances. It does not lead to the conclusion of Searle's "*Bad Argument*" by denying premise 4: during a hallucinatory experience, which is a genuine visual experience, there is not a "something" such as a sense-datum that is seen. Instead, there is a visual experience that is modified in a certain way. Thus the use of adverbs is appropriate to emphasize that the phenomenal character refers to a way of experiencing rather than to an object of experience. Instead of saying that I hallucinate something red, adverbialists rather say that I am hallucinating "redly". The problem with this strategy is that it faces difficulties for complex experiences. This is what Jackson calls the "*many properties problem*":

> *(1) I have a red, round after-image.*
> *is analysed as*

(2) I sense red-ly and round-ly
Hence,
(3) I have a red, square after-image and a green, round after-image.
is analysed as
(4) I sense red-ly and square-ly and green-ly and round-ly.
But (4) entails (2), while (3) does not entail (1); and so the conjunctive answer fails. In essence the point is that we must be able to distinguish the statements: 'I have a red and a square after-image', and 'I have a red, square after-image', and Ducasse does not appear to be able to do this. (Jackson 1975)

Jackson's point is that it is difficult to account for the structure of experience without objects of perception. This is because the idea of a structure seems to presuppose that of objects and relations between objects. How am I gonna say that the redness that I experience when looking at the apple is "inside" the experience of roundness and "in the middle" of a whole visual field "filled by" other colors and shapes? This lack of clearly analyzable structure that appears as a weakness of adverbialism at first glance could in fact be a strength.

Talking about "phenomenal redness" as if it was separated from the whole visual field in which it appears is at the root of a lot of problems. A phenomenal experience is *holistic*: its individual contents such as redness and painfulness are always parts of larger wholes from which they cannot completely be abstracted .There is not really an atomic red "sense-datum" inside my experience. As Husserl (1958, p. 71) puts it, consciousness is "*not something like a mere box in which things given simply are*". This point was also made by Carnap:

Modern psychological research has confirmed more and more that, in the various sense modalities, the total impression is epistemically primary, and that the so-called individual sensations are derived only through abstractions, even though one says afterward that the perception is "composed" of them: the chord is more fundamental than the individual tones, the impression of the total visual field is more fundamental than the details in it, and again the individual shapes in the visual field are more fundamental than the colored visual field places, out of which they are "composed". These psychological investigations have frequently been undertaken in connection with Gestalt theory. (Carnap 1927, p. 109)

"*Gestalt theory*" is a school of psychology that emerged in the early twentieth century in Austria and Germany as a theory of perception (Köhler 1967). One of the central claims of the theory was *phenomenal holism,* the view that wholes are prior to their parts in the realm of experience (Dainton 2010 ; Chudnoff 2013). An example of a visual whole that is prior to its individual parts, called a "gestalt"[7], is shown on *Figure 1*:

Figure 1. Example of a visual gestalt (left) and one of its parts (right)

When looking at the little "pie" on the bottom right of the gestalt, one has a different visual experience than when looking at the individual pie on the right. As Elijah Chudnoff writes:

> *There is a phenomenal difference between these two visual experiences. The difference is in their phenomenal content. Your visual experience of the bottom left pie represents it as a disc with a wedge that is occluded by a white triangle. Your visual experience of the isolated pie represents it as a disc with a wedge that is cut out. So your visual experience of the bottom left pie has a phenomenal character that distinguishes it from your visual experience of the isolated pie.* [...]
> *So we can conclude that your visual experience of the bottom left pie depends for its existence on the whole visual experience of which it is a part.* (Chudnoff 2013)

Of course, such an example does not show that phenomenal holism is true in a strong sense, which would be the thesis that every experience for a subject at a time is a sort of gestalt. Nevertheless, the fact that experience can be holistic in the sense pointed out by Gestalt theorists constitutes an obstacle for any theory that reifies appearances.

The intuition underlying adverbialism is that a conscious experience, instead of being comparable to a contentful box, should rather be compared to a vibrating string (Gert 2020).

[7] "Gestalt" means "form" in German.

A string vibrates in a certain way, and there is no vibration that is separated from the string itself. In a similar way, phenomenal redness is not an additional "thing" added inside a visual experience. Instead, it is more appropriately interpreted as the modification of an already existing "field" of consciousness. As the phenomenologist Merleau-Ponty writes:

> *Le « quelque chose » perceptif est toujours au milieu d'autre chose, il fait toujours partie d'un « champ ». [...] La pure impression n'est donc pas seulement introuvable, mais imperceptible et donc impensable comme moment de la perception.* (Merleau-Ponty 1945, p. 10)

This is not to say that experiences are literally mental fields that can be perturbed. The point is that the individual qualitative aspects of experience (such as colors and shapes for a visual experience) are "superposed" in such a way that each one of them is subjectively presented in a unique manner that is not abstractable from the whole. Adverbialism captures this feature greatly by allowing adverbs to be combined, as explained by Sellars:

> *Jackson fails to consider the possibility that the adverbialist might attempt to distinguish between*
> *I sense redly and squarely and*
> *I sense (red and square)ly*
> *The latter rather than the former being counterpart of*
> *I have a red, square after-image*
> (Sellars 1975)

What is particularly interesting with adverbialism is that it highlights how apparently simple aspects of phenomenal experience such as colors are in reality always parts of a very complex whole. Nobody has ever experienced a pure quale of red, one rather experiences a "red and square" quale that is itself part of a more complex qualitative experience. Both the red and the square would appear differently if they were not visually combined the particular way they are. This is for instance what the famous painter Kandinsky writes:

> *If two circles are drawn and painted respectively yellow and blue, brief concentration will reveal in the yellow a spreading movement out from the centre, and a noticeable approach to the spectator. The blue, on the other hand, moves in upon itself, like a*

> *snail retreating into its shell, and draws away from the spectator.* (Kandinsky 1911, pp. 80-81)

According to Kandinsky, a visual experience of a circle is not the same whether it is painted in yellow or in blue. In other words, roundness presents itself differently when it is sensed blue-ly and when it is sensed yellow-ly. That said, it is undeniable that there is something shape-wise similar between an experience of a yellow circle and that of a blue circle. In both cases, a normal human being is conscious of a circle. Being conscious of a circle *as a circle* necessitates the concept of a circle. When we see a blue circle, our concept of circle is activated, we have thoughts about the circle that are somehow integrated into the visual experience. Thus it is impossible for us to know what is a first-order "raw" experience of roundness devoid of conceptual meaning. Phenomenology is holistic also in the sense that first-order raw phenomenology is never separated from conceptual content in conscious experience, which makes the analysis of the latter particularly difficult.

A good example showing how conceptual content shapes visual experience is the famous duck-rabbit illusion. On *Figure 2*, the same visual information can either be interpreted as a duck or as a rabbit. On the one hand, it is hard for the sense-data theory to account for this illusion because it is not clear whether or not the duck percept presents the same sense-data as the rabbit percept. On the other hand, Russelian representationalist theory of perception such as Dretske's or Tye's would lack the concepts to fully describe what happens. What is remarkable with this illusion is not merely that the perceiver alternatively modifies its representational state, but that what can be seen as shape-wise and color-wise invariant is literally shaped by conceptual content. In other words, the visual mode of presentation stays the same while the represented properties are changing. Thinking about a rabbit or a duck changes the visual content that is presented in experience, a phenomenon that can only be described by conceptually distinguishing what is represented from how it is presented.

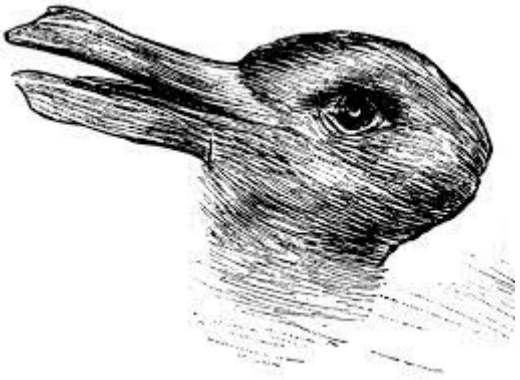

Figure 2. The duck-rabbit illusion

How is all of this relevant for the problem of artificial qualia? In part 1.7, I discussed the multiple realization thesis: the same phenomenal property can be realized by different physical properties. For example, the redness of my experience right now is realized by a neural activation in my visual cortex, but the same property could in principle be realized in the brain of a honey bee by a very different process. The problem is that, if phenomenal redness is to be understood as an adverbial modification of experience, the question of its realization becomes much more difficult. This is because the process that realizes my experience of red right now is the one that realizes my conscious experience as a whole. There is no sub-process that realizes the redness of my experience individually because there is no such thing as a pure individual experience of red or a red sense-datum. Instead of seeing a red sense-datum, I argued that it is more appropriate to say that I experience (red and square)-ly or (red and round)-ly. Phenomenal experiences of squares and rounds are part of my experience of redness. Therefore, even if neuroscientists identify a neural correlate of my experience of redness in particular, it is not possible to interpret this neural state as the authentic realizer of phenomenal redness. As Searle explains:

> *Most of the discussions I have seen of the NCC* [neural correlates of consciousness] *are confused because the researchers are looking for an NCC for a particular element of the conscious field, such as, for example, the experience of the color red. But that experience occurs in a subject who is already conscious. So the NCC could not possibly give us sufficient conditions for consciousness because the subject has to be already conscious in order that the NCC in question can cause a particular perceptual experience. The basic insight is this: We should not think of perception as*

> *creating consciousness, but as modifying the preexisting conscious field.* (Searle 2004)

So how exactly are we going to understand conscious experiences? In the next part, I will look at this mysterious conscious field which corresponds to what I call a quale, and see if it can possibly be rigorously modeled.

2.5. The structure of qualia

Interestingly, the concept of qualia that I am defending in this work has interesting similarities with its first use by C.S. Peirce:

> *The quale-consciousness is not confined to simple sensations. There is a peculiar quale to purple, though it be only a mixture of red and blue. There is a distinctive quale to every combination of sensations so far as it is really synthesized - a distinctive quale to every work of art - a distinctive quale to this moment as it is to me - a distinctive quale to everyday and every week - a peculiar quale to my whole personal consciousness. I appeal to your introspection to bear me out in this.* (Peirce 1866)

Interpreting the "*whole personal consciousness*" as a quale highlights the holistic character of qualia mentioned in the previous part. There is one experiential process that we can call quale-consciousness. If the goal is to build an artificial system that experiences the world, then it must be found how the system could somehow produce its own quale-consciousness. Of course, the problem is that we have no idea what quale-consciousness is. What we know however is that we, humans, are quale-conscious. Thus we have no choice but to start from our own experience. We must understand how experience is subjectively lived. From this perspective, modeling consciousness means modeling the phenomenology itself. The idea of starting from experience as it is subjectively lived may be linked to the project of phenomenologists such as Husserl and Merleau-Ponty, as discussed in part 1.6. The important difference is that phenomenologists were not interested at all in the realization or generation of experience and they may have found this problem absurd. Nonetheless, the approach of the phenomenologists is, at least in its spirit, essential for a full comprehension of consciousness. It is necessary to understand consciousness as it is experienced because the explanandum of a

theory of consciousness is the subjectively lived experience itself, the quale-consciousness that all of us *are*.

Saying that I am myself an experiential process, or a quale-consciousness, means that at any given moment, I introspectively find nothing else but sensations and thoughts in my experience. This idea has been notably defended by Hume:

> *For my part, when I enter most intimately into what I call* myself*, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe any thing but the perception.* (Hume 1740, p.165)

According to Hume's "*bundle*" theory of the self, the mind "*is nothing but a heap or collection of different perceptions, united together by certain relations, and suppos'd, tho' falsly, to be endow'd with a perfect simplicity and identity"* (ibid., p. 137). As argued in the previous part, interpreting the phenomenal content of experience as a collection of little things is misleading. Instead, I argued that a better metaphor was that of the "field". It is for instance used by Searle in the following passage:

> *Conscious states are also subjective in the sense that they only exist as experienced by some human or animal subject; and in nonpathological cases, they always come to us as part of a unified conscious field. That is, we don't just have the qualia of the taste of coffee in our mouth, the slight headache, and the sight of the landscape out the window; rather we have all of these as part of a single unified conscious experience.* (Searle 2005)

Thinking of the quale-consciousness as a field emphasizes its unity (Bayne 2010 ; Brook & Raymont 2017). One may object that this unity is just an introspective illusion, as Hume argued. For example, David Rosenthal insists that there is only a "sense *of the unity of consciousness*" (Rosenthal 1986, p. 344). In other words, consciousness *appears* as unified but it may not be in reality. This objection does not hold because it is precisely the appearance itself that we try to explain with a theory of quale-consciousness. At any given time, all the phenomenal content of my visual experience is presented at once, in what *appears* as a visual *field*. And this is not restricted to visual experience. The conscious field is

the totality of what is presented to me in experience at a given time: thoughts, emotions, pain, sounds and odors.

This point is important because it shows that we will never understand what is the exact causal substrate of a quale of redness or pain if we do not first clearly understand the full underlying structure of qualitative experiences. The only possible method to reveal the structure of an experience of red is to somehow "reverse-engineer" human qualia. We must understand the structure of this field of consciousness that we are. To better understand what it means, let's suppose again that we want to create a robot that experiences colors. We do not want the robot to merely detect properties of external objects, we want it to experience objects red-ly and blue-ly in a similar way that we do. Is it possible to generate artificial color qualia without first figuring out how to make the robot able to feel pain, pleasure and emotions? At first glance, it seems that these are two very different problems. Feeling and seeing appear to be two independent conscious phenomena, so it seems intuitive to treat them separately. However, the ability to have a visual field full of colors may be linked to the ability to feel emotions in a way that is difficult to notice for normal perceivers, but that can be revealed thanks to anomalous cases. The neuroscientists Oliver Sacks and Robert Wasserman report the spectacular case of a painter that, after losing his ability to see colors because of a brain damage, gradually witnessed radical changes in his experience:

> *It was not just that colors were missing, but that what he did see had a distasteful, 'dirty' look, the whites glaring, yet discolored and off-white, the blacks cavernous - everything wrong, unnatural, stained, and impure.* [...]
> *He found foods disgusting in their grayish, dead appearance and had to close his eyes to eat. But this did not help very much, for the mental image of a tomato was as black as its appearance.* (Sacks & Wasserman, 1987)

This example suggests that there are yet unknown connections between the different qualitative aspects of our experiences. Whether or not the ability to have phenomenal experiences of color depends on emotional abilities should be an empirical question. The problem is that we lack the conceptual framework that could allow such questions to be addressed in a systematic way. We need to have models of qualia.

What I call models of qualia can be linked to the concept of *qualia space*:

> *We define qualia space Q to be the space of all possible conscious experiences. [...] The structure of qualia space allows us to consider and even answer in a precise way such questions as: Is there a continuous path from the sensation of blue to the sensation of pain?* (Stanley 1999)

A qualia space is an abstract representation of the qualities of experience where their structural relations are highlighted. It is related to what Austen Clark calls a "*quality space*" (Clark 2000). Examples of qualia spaces are color spaces (Kuehni 2003) and sound spaces (Casati & Dokic 1994). These are actually instances of qualia *sub*spaces that are meant to model one particular type of quality. These models are useful to understand the cognitive content of qualitative experiences, such as the way colors and sounds represent objective properties of the world. This is for instance what Paul Churchland proposes when he investigates the relationships between the color-qualia space and the reflectance-profile space (Churchland 2007). Although his analysis is undoubtedly valuable for a better understanding of visual perception, the "*structural homomorphisms*" that he identifies between the phenomenology of colors and the properties of surfaces cannot explain the nature of color experiences themselves, their purely presentational content.

Why do the reflectance properties of surfaces present themselves colorfully instead of say, auditorily? And why do the acoustic vibrations of guitar strings present themselves musically instead of colorfully? These questions might seem odd at first glance since it seems natural for most of us that reflectance properties are visual and that acoustic vibrations are auditory. This is a category mistake: there is nothing intrinsically visual to a surface and nothing intrinsically auditory to an acoustic vibration, in the same way that there is nothing intrinsically painful to a body damage. In part 1.1, I relied on a thought experiment to show that painfulness is a particular way of experiencing a body property, what I called a mode of presentation of a property in part 2.3. In the case of perceptual experiences, there actually exist concrete examples that can highlight anomalous perceptual modes of presentation. A first very interesting one is the case of *synesthesia*, a syndrome in which the conscious access to certain properties involves unusual qualia. Cytowic reports a case of a woman that "sees" sounds and describes her experience as follows:

> *The only real problem is that when I am driving and a very loud sound comes on such as loud music or the Alert Test tone and it is hard to see. The image intensity is directly proportional to the sound level. People laugh when I say, "turn that down, I can't see where I'm driving."* (Cytowic 1989, p. 51)

If acoustic vibrations can be presented visually, the nature of visual qualia cannot be completely understood as representations of surfaces. It is not sufficient to know how color qualia relate to the external world, we also need to understand the way visual phenomenology fits into the whole experiential field. The example of synesthesia suggests that there exists a relationship between the phenomenal color space and the phenomenal sound space. Are there more fundamental ways of experiencing the world that structure both the visual and the auditory phenomenology at a deeper level? Both auditory and visual experiences are spatial, thus it makes sense to talk about *spatial qualia*. As Pete Mandik writes:

> *I have encountered people who thought it weird to say that there were spatial qualia. They granted that there were color qualia and odor qualia, but not space qualia. As I see it, if there are any qualia at all, then it is very likely that there are spatial qualia. It is for instance very likely that spatial qualia are underlying both conscious vision and conscious hearing.* (Mandik 1999).

In order to understand what a spatial quale could be, another concrete example happens to be illuminating. More than half a century ago, the neuroscientist Paul Bach-y-Rita and his colleagues developed a *Tactile Vision Substitution System* that they presented as "*a practical aid for the blind and as a means of studying the processing of afferent information in the central nervous system"* (Bach-y-Rita & al. 1969). Their idea was to enable blind people to "see" with their tactile capacities. Blind people learned to associate tactile vibrations with the content of images captured by a camera. The images were pixelated and then the pixels were transformed into vibrating patterns that the blind subjects could feel on the skin of their back. Thanks to this device, the perceptual system of the subjects quickly learns to transform the stimuli received through the skin into a real perception of space, even for people that were born blind. Numerous studies after the pioneer work of Bach-y-Rita have proven the effectiveness of this substitution of vision by tactile stimuli, particularly for the recognition of simple shapes. Would it therefore be correct to claim that blind people have *visual* experiences when using the device? Bach-y-Rita's answer is that "*although the early system*

was termed a tactile visual substitution system, we have been reluctant to suggest that blind users of the device are actually seeing" (Bach-y-Rita & Hughes 1985). On the other hand, Michael Morgan claimed that the blind patients are really seeing with the device (Morgan 1977). He argued that the structural nature of the perceptual system does not offer any criteria for distinguishing seeing from not seeing. So are they seeing or not? If the blind subjects are able to consciously experience a shape through the device, it is probable that there is something it is like for them to experience space. They are able to realize a quale of roundness or a quale of squareness. It is in this sense that we can talk about spatial qualia. Whether or not their spatial experience should be labeled "visual" is mainly a verbal dispute. It is equivalent to asking if a bat has a visual experience when using echolocation to perceive its environment. The experience of the bat and the experience of the blind patients are "visual" only in the sense that they share spatial aspects but probably not in the sense that they involve color aspects.

Bach-y-Rita later claimed that the crucial difference between normal visual experiences and the experiences enabled by his device was that the latter failed to generate qualia:

> *Les sujets entraînés à utiliser le système de substitution ont remarqué l'absence de qualia, qui peut souvent être très perturbante. Ainsi, des sujets bien entraînés sont-ils profondément déçus quand ils explorent le visage de leur femme ou de leur petite amie et découvrent que, même s'ils peuvent en décrire les détails, l'image ne possède aucun contenu émotionnel.* (Bach-y-Rita 1997)

Interestingly, he uses the term qualia to refer specifically to the affective dimension of experience. He seems to suggest that an experience is qualitative only when it has a *value*. Does it mean that the ability to see colorfully is a kind of affective capacity? It is not clear at all that such a conclusion can be drawn since Bach-y-Rita also mentions the case of blind patients who recover their sight thanks to an operation, and for whom colors have no affective qualities (Gregory & Wallace 1963). It is not easy to decide whether or not such patients see colors in the phenomenal sense. They may be able to discriminate between red and green without having the typical "what-it's-like-ness" that normal sighters have when they see colors. Again, a "visual" experience seems to involve spatiality in a more fundamental way than colorfulness. This suggests that qualia are *hierarchically structured*: if we reverse-engineer human qualia we may probably find that there are fundamental ways our

experiences are formed, and spatiality and raw affective sensations are arguably ones of them. I will get back to this in the third part of this work.

To better understand how we could possibly understand the structure of qualia, let's look at a very interesting contemporary scientific attempt to model phenomenology: the integrated information theory (IIT) of consciousness (Tononi & Koch 2015 ; Tononi & al 2016). Tononi's original insight was that, in order to solve the Hard problem of consciousness, we should start from first-person experience and then find the mechanisms in the world to which they correspond (Tononi 2004). This is very close to what I presented as methodological dualism in part 1.6. Tononi is a neuroscientist and he is convinced that experiences are identical to mechanisms happening in the brain. Thus, his goal is to *map* qualia to brain processes. For example, IIT postulates that "*consciousness is unified: each experience is irreducible to non-interdependent subsets of phenomenal distinctions*" which is a phenomenological axiom that is close to what others call the unified conscious field, as mentioned earlier. From that, it is inferred that "*the cause–effect structure specified by the system must be unified: it must be intrinsically irreducible to that specified by non-interdependent sub-systems*" (Tononi & Koch 2015). The idea is that the brain realizes an integrated process that *is* an experiential process.

Tononi and his colleagues formalized a rigorous mathematical framework that can supposedly account for the underlying mechanisms of qualia, and this is essentially what the IIT is. The advocates of the IIT claim that the theory makes it possible, at least in principle, to deduce what an experience is like from brain activity alone. For example, Naotsugu Tsuchiya explicitly wrote that the IIT has the potential to answer Nagel's question "*what is it like to be a bat?*":

> *Why a particular sense, such as vision, has to feel like vision, but not like audition, is totally puzzling. This is especially so given that any conscious experience is supported by neuronal activity. Activity of a single neuron appears fairly uniform across modalities and even similar to those for non-conscious processing. Without any explanation on why a particular sense has to feel the way it does, researchers cannot approach the question of the bats' experience. Is there any theory that gives us a hope for such explanation? Currently, probably none, except for one. Integrated*

information theory (IIT) has potential to offer a plausible explanation. (Tsuchiya 2017)

Tsuchiya's paper outlines a schema that summarizes pretty well the strategy of the IIT, as shown on *Figure 3*:

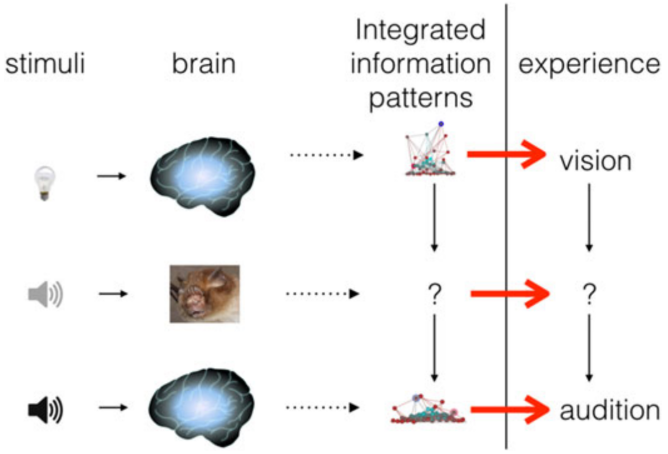

Figure 3. Schema representing how IIT addresses the question of bat's phenomenology (Tsuchiya 2017)

The essential point to understand here is that the advocates of IIT claim that there exist isomorphisms (structural identities) between qualia and brain structures. These isomorphisms can be modeled mathematically in an abstract qualia space (Q):

> *We recognize intuitively that the way we perceive taste, smell, and maybe color, is organized phenomenologically in a "categorical" manner, quite different from, say, the "topographical" manner in which we perceive space in vision, audition, or touch. According to the IIT, these hard to articulate phenomenological differences correspond to different basic sub-shapes in Q, such as grid-like structures and pyramid-like structures. In turn, these emerge naturally from the underlying neuroanatomy and neuronal activity patterns.* (Balduzzi & Tononi, 2009)

According to the IIT, a quale is a brain process. The spatiality of a quale corresponds to particular structures of neural networks. But isn't it odd to claim that an experience of space is a process happening inside the head? This would mean that when I look at a man walking

in the street, the experience of the environment that the man is having is a process that is completely taking place inside the boundaries of his skull. It seems more natural to believe that his experience of space is a process that involves real space out there, at least in normal situations. Or is it?

This question will be addressed in the third part of this work. I will go back to the roots of contemporary cognitive science, whose origin is common with computer science and artificial intelligence. The idea that experience is a process happening inside the head will be challenged. On the other hand, I will show that we face other important issues when we claim that an experiential process includes the environment of the experiencing subject. Then I will discuss the evolutionary origins of experience to see if we cannot explain experience with biology. Finally, I will get back to the *Hard problem* of consciousness: it may not be possible to clearly distinguish between experiential and non-experiential processes and this would imply important metaphysical consequences.

3. "Artificial" qualia?

3.1. Qualia and intentionality

In a 1950 article, Turing famously asked the question *"Can machines think ?"* (Turing 1950). Turing realized that the answer to this question largely depends on the definition of "thinking". Hence, he proposed to remove the ambiguity by presenting a procedure that could directly test the thinking abilities of a machine. The test was called the "*imitation game*": a human interrogator talks with two interlocutors through a chat interface where one is a human and the other is a machine, and the interrogator's role is to identify which one is the machine. Thus the goal of the machine is to imitate typical human answers. Turing conjectured that if the imitation of the machine was sufficiently good for the interrogator to be wrong most of the time, it would be legitimate to claim that the machine is able to think. Contrary to the word "think", what Turing meant by "machine" has a precise definition that he formalized and is thus referred to as the *Turing machine* (De Mol 2019). A Turing machine is an abstract model of a *computational process*. While a Turing machine is an abstract mathematical formalism, a computational process is a concrete process happening in the world. A typical example of a computational process is what happens in a digital computer such as laptops and

smartphones. A laptop can abusively be called a Turing machine in the sense that everything that it does can be abstractly described as a sequence of operations on *symbols* such as '0' and '1'. This is a powerful concept because it means that, in principle, the same computational process that is realized by electrical currents in a laptop could in principle be executed by a completely different artifact. Turing's idea was revolutionary as it was going to be at the foundation of both artificial intelligence (AI) and cognitive science (McCarthy & al. 1955). What he suggested in his pioneer article is that thinking is essentially a computational process realized by the human brain, and therefore an abstract description of the thinking process should in principle allow us to reproduce it artificially.

The thesis that thinking is a computational process is often called *computationalism*, and it has been the orthodox view in the philosophy of mind of the 1960's and the 1970's (Rescorla 2020). In 1980, Searle proposed an influential argument against computationalism, based on what is known as the *Chinese Room* thought experiment. In his article, Searle invites us to imagine a man, locked alone inside a room, with a huge instruction book in which are written, in English, a lot of rules to answer a question written in Chinese. The man in the room cannot read a word of Chinese but if he receives from the outside a paper with a question written in Chinese, he can always find the correct rules that allow him to write Chinese pictograms that correspond to a typical answer to the question. The man in the room behaves exactly like a Turing machine: he manipulates symbols (the Chinese pictograms) and performs operations that are specified by a set of rules that he just follows blindly. From the outside of the room, it appears as if the man understands Chinese. However, nothing in the process really corresponds to what we usually call "understanding". If the person in the room does not understand anything of the manipulated symbol, it is difficult to maintain that "*somehow the conjunction of that person and bits of paper might understand Chinese*" (Searle 1980). The thought experiment was meant to emphasize an important difference between the thinking process that is realized by human beings when they answer a question on the one hand, and an imitation process that can in principle be performed by any sufficiently complex Turing machine on the other hand. According to Searle, what makes the difference between the two is that a genuine thinking process is *intentional*.

Intentionality is the power of minds and mental states to be *about* things, properties and states of affairs (Jacob 2019). When the man in the room sees the word "cat" written in English, he *understands* it in the sense that he undergoes an *experience* that is *about* cats. By contrast,

when the man receives Chinese pictograms that correspond to the question "Are you a cat ?", and answers with the pictogram corresponding to the word "No" thanks to his instruction book, there is nothing in this process that is really about cats. Of course, it could be objected that there is in fact a Chinese pictogram that is about cats. However, the "aboutness" of a pictogram is only *derived* from its use by Chinese people: if all the Chinese speakers of the world were to disappear, the Chinese pictograms would not be about anything anymore. In other words, the *primary* source of intentionality lies in the mental states of the Chinese speakers, hence Chinese pictograms inherit their intentional properties from human thinking. This immediately raises a question: where does this mysterious primary intentionality of human thinking come from?

The difference between the real thinking process of the man in the Chinese room when he receives a question in English and the computational process performed when he receives a question in Chinese can be expressed in terms of qualia. When an English speaker sees a question written in English, there is something it is like for him to understand the question. He has a subjective experience of the understanding of the word "cat" that he does not have when seeing the Chinese "cat" pictogram. This subjective experience corresponds to what Galen Strawson calls an "*understanding-experience*":

> *Does the difference between Jacques (a monoglot Frenchman) and Jack (a monoglot Englishman), as they listen to the news in French, really consist in the Frenchman's having a different experience? [...]*
> *It is certainly true that Jacques's experience when listening to the news is very different from Jack's. And the difference between the two can be expressed by saying that Jacques, when exposed to the stream of sound, has what one may perfectly well call 'an experience (as) of understanding' or 'an understanding-experience', while Jack does not.* (Strawson 1994, chap. 1.4)

In the same way, for a monoglot Englishman, the experience of seeing the word "cat" written in English is very different from that of seeing the corresponding Chinese pictogram. Obviously, the difference in this case is visual. Nonetheless, the two scenarios are also different in another very important respect: for an English speaker, reading an English word involves a *cognitive quale* that is not realized when he looks at a Chinese pictogram. An understanding-experience is a cognitive quale in the same way that a painful experience is a

pain quale. It is in virtue of the understanding-experience that there is cat "aboutness" when the man looks at the word written in English. From this perspective, the problem of intentionality is a special case of the more general problem of qualia.

The view that intentionality is grounded in qualia is called the *phenomenal intentionality thesis* (*PIT*) (Bourget & Mendelovici 2019). According to a strong version of the *PIT*, intentionality entirely comes from phenomenal consciousness. For example, Mendelovici claims that "*intentionality is simply identical to phenomenal consciousness*" (Mendelovici 2018, p. XV). This strong version of the *PIT* is probably too radical. Unconscious mental representations such as unconscious beliefs are intentional without being phenomenal. According to a moderate version of the *PIT* defended by Searle, it is only because of the *disposition* of a representation to be experienced that it is genuinely about something. This point is close to the discussion in part 2.2 about the distinction between mental and non-mental representational states.

To understand why, let's take another look at the example of thermostats. A thermostat represents the ambient temperature of a room thanks to a physically realized internal state. Should we say that the internal state of the thermostat is *about* the ambient temperature? In a sense, it obviously is, since by definition a representational state is about what is represented. One way to argue that the thermostat has a representational state about a temperature is to point out that it contains information about the ambient temperature. This is so because there is a lawful dependence between the state of the thermostat and the ambient temperature of the room: the state of the thermostat is systematically correlated to the ambient temperature. The problem is that the state of any other macroscopic object in the room, such as a table, is also lawfully dependent on the ambient temperature and thus also contains information about the ambient temperature. It could be objected that a table is not a cognitive system like the thermostat. A thermostat, contrary to a table, *uses* the information about the temperature to do something. In this case, to regulate the temperature of the room. But this notion of "use" of information is just a way of interpreting a causal process that has nothing "cognitive" in and of itself. An ice cube can be said to "use the information" about the ambient temperature to melt, but obviously an ice cube is not a cognitive system, at least not in any interesting sense. The point is that, in the physical world, there are just continuous *causal* processes and we interpret some of them as cognitive or informational because it is useful to do so in certain contexts.

Now, *mental* processes are cognitive and informational in a significantly different sense. This is because, as argued in part 2.3, the information contained in a mental representation can be presented in a certain way, it can be like something to access it. Searle calls this the "*aspectual shape*" of intentional states:

> *Noticing the perspectival character of conscious experience is a good way to remind ourselves that* all intentionality is aspectual. *Seeing an object from a point of view, for example, is seeing it under certain aspects and not others. In this sense, all seeing is "seeing as." And what goes for seeing goes for all forms of intentionality, conscious and unconscious. All representations represent their objects, or other conditions of satisfaction, under aspects. Every intentional state has what I call an* aspectual shape*.* (Searle 1992, chap. 6)

This corresponds to what I called this the mode of presentation of a representational content. I argued in part 2.3 that this is exactly what the phenomenal content of an experience is. When Mary is in pain for the first time, she acquires information *about* her body *from* her phenomenal experience of pain. By contrast, a robot can acquire information about the damage of its artificial body in the same sense that a thermostat acquires information about the ambient temperature. In both cases, it is a way we interpret a causal process that is not mental and thus not intentional. In this view, only an informed experiential process is truly about something. In a world devoid of experience, there are just causal and other physical dependencies between objects but there is no intentionality. This is why Brentano, who introduced the concept in modern philosophy, viewed intentionality as "*the mark of the mental*" (Crane 1998). According to Brentano, intentionality "*is characteristic exclusively of mental phenomena*" and *"no physical phenomenon manifests anything like it*" (Brentano 1874).

Let's now come back to the Chinese room. How is it possible that an experience of reading something in English is a cognitive quale in a way that the process happening in the Chinese room is not? A lot of answers to Searle's argument rely on the idea that the room itself somehow "understands" Chinese. Ray Kurzweil argues that Searle contradicts himself when he claims that the system composed of the man and his book can speak Chinese without understanding it (Kurzweil 2002). This was in fact Turing's original point: in a situation

where a system behaves exactly as if it “thinks”, it makes no sense to ask whether or not it “really” thinks. We will never have a consensual definition of “thinking”, so in order to avoid verbal dispute it makes sense to replace the question by a practical test. The underlying idea is that when it comes to thinking, there is no difference between the real process and its simulation. It would be absurd to ask someone to simulate the resolution of a problem without “really” solving it. Hence, if a digital computer or any other system such as Searle’s Chinese room is able to simulate the understanding of Chinese, it is legitimate to claim that the system understands Chinese.

Is the same argument applicable to qualia? One could argue that everything happens as if the Chinese room itself is “experiencing” the understanding of Chinese pictograms. We will never have a consensual definition of “experience”, so if the Chinese room can be said to “think”, why not bite the bullet and also claim that it can have a conscious experience? The reason is that in the case of experience, the conflation between the real process and its simulation does not hold anymore. Mary can act “as if” she feels pain without really feeling it and there is no particular challenge in building an artefact that exhibits pain-like behavior, as already mentioned in part 1.7. That said, the idea that the system consisting of the man and his book could be conscious as a whole is not that absurd and is linked to the *extended mind* thesis (Clark & Chalmers 1998). After all, the man could, at least in principle, memorize all the rules of the book and thus answer questions in Chinese all by himself, as Searle points out in his article. From the outside, it would appear as if the man is a Chinese speaker. However, from the man’s point of view, there will be an important difference between the experience of applying the memorized rules to answer a Chinese question and the experience of understanding a message written in English. The difference is phenomenological and hence supports my point: there is no way to account for *conscious* understanding without a concept of cognitive qualia. The latter are arguably the most complex type of qualia. Thus, before tackling the problem of human cognitive qualia, it makes more sense to address in priority the general question of qualia. The challenge of artificial qualia is very different from Turing’s original problem of machine thinking. Can machines feel? There is no reason to believe that the implementation of a computer program that simulates thinking processes has anything to do with feelings.

> *Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality.* (Searle 1980)

So what is so special about the brain?

3.2. Artificial brains and functionalism

It is important to note that Searle's Chinese room argument was targeting a restricted class of computational processes, as it was based on the dominant paradigm of the time. It corresponds to what has been later called "*Good Old Fashioned AI*" or "*symbolic AI*" (Haugeland 1989), namely the methods based on high-level human-readable symbols and explicit rules operating on these symbols. What the brain does is obviously very different from the symbol manipulation performed by the man in the Chinese room, so the thought experiment does not undermine computationalism at first glance. It can still be considered that a computational process that precisely replicates the activity of the brain would involve an experience. Let's take a closer look at this hypothesis. What does it mean for a computational process to replicate the activity of the brain? As explained in the previous part, a computational process is a physically realized process that can be described formally. As Chalmers (2011) puts it "*a physical system implements a given computation when the causal structure of the physical system mirrors the formal structure of the computation*". Thus, for the sake of clarity, we can forget about the notion of computation and focus on that of *causal process*. To put it in the simplest possible way, our problem of interest in this work can be formulated as follows: what exactly are the causal processes that involve qualia?

Let's assume for now that the causal substrate of qualia is to be found in the brain. In other words, there are some causal processes happening in the brain that are sufficient for qualia to "arise". If we were able to build an artefact that reproduces the causal powers of the brain, this artefact would supposedly also generate qualia. Let's call this artefact an artificial brain. We have no idea yet what artefacts could be artificial brains but we can imagine a procedure that would allow us to go from a real brain to an artificial one, as Chalmers explains :

> *We can imagine, for instance, replacing a certain number of my neurons by silicon chips. In the first such case, only a single neuron is replaced. Its replacement is a*

> *silicon chip that performs precisely the same local function as the neuron. We can imagine that it is equipped with tiny transducers that take in electrical signals and chemical ions and transforms these into a digital signal upon which the chip computes, with the result converted into the appropriate electrical and chemical outputs. As long as the chip has the right input/output function, the replacement will make no difference to the functional organization of the system.* (Chalmers 1995b)

The idea is that, under the assumption that the causal role of one neuron can be assured by an artificial object such as a silicon chip, the possibility of a whole artificial brain made of silicon is entailed by recurrence. This thought experiment is meant to support the idea that qualia are "*organizational invariants"*: "*when experience arises from a physical system, it does so in virtue of the system's functional organization*" (ibid.) The principle of organizational invariance is central to the thesis called *functionalism* (Levin 2018). Presented in this way, functionalism does not seem to be an unreasonable view. Why not suppose that a silicon brain can produce qualia? The problem is that this principle of organizational invariance leads to a slippery slope.

If we accept that a silicon brain can produce qualia, why not also accept that a giant artificial brain made of water pipes can also generate experience? (Searle 1980) After all, what a silicon chip does is nothing more than transmitting an electrical current. If only the organization of the system matters, it should not be important whether the current is realized by water or by electricity. So, of course, the replacement procedure of neurons by water pipes is not possible. However, once we accept that the causal process realized by a silicon brain is able to produce qualia, it is difficult to understand why, at least in principle, a very sophisticated plumbing machine could not be conscious. Block (1978) goes even further in an article criticizing functionalism. He argues that if we asked the whole population of China to organize as if each individual is a neuron, where the synaptic interactions are replaced by radio communications, the Chinese nation would become a giant artificial brain. Eric Schwitzgebel makes a very similar point in his article titled "*If Materialism Is True, the United States Is Probably Conscious*":

> *What is it about brains, as hunks of matter, that makes them special enough to give rise to consciousness? Looking in broad strokes at the types of things materialists tend to say in answer– things like sophisticated information processing and flexible,*

> *goal-directed environmental responsiveness, things like representation, self-representation, multiply-ordered layers of self-monitoring and information-seeking self-regulation, rich functional roles, and a content-giving historical embeddedness – it seems like the United States has all those same features. In fact, it seems to have them in a greater degree than do some beings, like rabbits, that we ordinarily regard as conscious.* (Schwitzgebel 2015)

This is also one of the main problems of the integrated information theory (IIT) presented in part 2.5. Tononi assumes that experiences are generated in the brain. Then he develops an abstract model of the processes that realize qualia, what he calls a qualia space. Once the process is abstracted, he infers that any physical process that realizes the same structure will realize the same experience. He makes this inference because of the principle of organizational invariance: it is because of the way neurons are organized that qualia arise, not because of the physical neurons themselves. Finally, to be consistent with his own theory, he has no choice but to be committed to the view that any mechanism can be conscious. This is why the IIT is often interpreted as *panpsychist*[8] theory: consciousness is "*here, there and everywhere*" (Tononi & Koch 2015). Scott Aaronson demonstrated that the IIT predicts that an 2D grid of several billions of identical logical gates can be as conscious as a human being without doing anything (Aaronson 2016). Tononi (2014) responded that it is not such an implausible claim. Therefore, he would also claim that a giant plumbing machine can in principle have the same experiences as that of humans to save his theory. Are we rationally bound to believe that any system that is organized in the same way as the brain can have conscious experiences?

It is of course always possible to bite the bullet and accept that, at least in principle, a system made of water pipes is able to feel pain. However, apart from being very counterintuitive, this claim has no empirical support. The only reason to accept it would be a strong commitment to functionalism. The problem is that, when refusing functionalism, we are brought back to our initial question: what is so special about the brain? As Dennett rightfully argues, denying functionalism seems to entail that biological brains have mysterious powers, which would imply a form of vitalism:

[8] I will get back to panpsychism in part 3.5.

> *Supposing, then, that a manufactured biochemical duplicate* [of a brain] *would feel pain, [...], what difference could it make if we use other materials? Only two replies, both insupportable, occur to me: (1) organic compounds are capable of realizing functional structures with capacities of a sophistication or power in principle unrealizable in non-organic materials, or (2) though an inorganic replica might succeed in duplicating a human being's functional structure, the states in it functionally isomorphic to human pain states would fail to be genuine pain states because the biochemistry of pain state realizations is essential. These are both highly implausible vitalistic claims, and any skeptic led to defend his view in this territory has simply been led astray.* (Dennett 1978)

In part 3.4, I will discuss the relationship between qualia and life. Before that, let's come back to this initial supposition that a biological brain is somehow able to "generate" qualia.

3.3. The sensorimotor approach to experience

As discussed previously, qualia are ways real properties of the world are consciously accessed. In most normal situations, a pain experience is a particular way of accessing body damage and a color experience is a particular way of accessing complex properties of external surfaces. From this perspective, the idea that consciousness is produced by a mechanism of the brain is odd. Indeed, if consciousness is an access to real things out there, it seems natural to include these real things in the process by which they are consciously accessed. For example, when I see a red apple, my qualitative experience of the red apple is the way this real apple is presented to me. Wouldn't it make sense to claim that the real apple is part of the process I call the experience of the apple? Why suppose that my brain is solely responsible for experience? The assumption that brains have the power to generate qualia may have been flawed in the first place. This is what the psychologist Kevin O'Regan claims:

> *In my statement: "Visual experience is not generated in the brain," the important point is not the word brain, but the word generated. Visual experience is simply not generated at all. Experience is not the end product of some kind of neural processing. I think the comparison with "life" is useful. Where is life generated in a human? Is it generated in the heart? In the brain? In the DNA? The question is meaningless, because life is not generated at all. It is not a substance that can be generated; it is*

> *nothing more than a word used to describe a particular way that an organism interacts with its environment.* (O'Regan 2011, p. 65)

This is close to what was argued in part 1.5: as for life, qualitative experiences are more appropriately interpreted as dynamical processes than as property-bearing substances. If qualia are processes, they obviously include brain activities, but they may also involve the actions of the body and its interaction with its environment. The view that conscious experiences necessitate the action of the body is often referred to as the *sensorimotor approach* to experience (O'Regan & Noë 2001). It is very close to what is known as the *enactive* or *embodied* approach to cognition (Varela & al. 1993) to which I will come back in the next part.

A relevant example to emphasize the role of the body in conscious experience is the case of the feeling of the "*softness of a sponge*" (O'Regan 2011, p. 294). When squeezing a sponge, there is a particular way its softness is presented in experience. Hence, there is a softness quale. Is this quale generated in the brain? It seems more reasonable to believe that it is the interaction between the real sponge and the hand of a conscious subject that gives rise to this soft quality. In this view, an isolated computing machine could never feel what it's like to squeeze a sponge. What is needed to give rise to an artificial quale of softness is a moving *robot* that possesses an artificial brain and an artificial body, with the correct sensorimotor abilities to interact with a sponge.

According to the advocates of the sensorimotor approach, what is true for softness is true for every experience. Intuitively, the role of the body is not obvious at all in sensory experiences such as vision. This is because vision is usually thought of as a passive experience: one opens the eyes and receives information from the outside that is somehow transformed by the brain into a visual field. However, as discussed in the second part, phenomenal colors cannot simply be interpreted as properties of any "things" in the straightforward sense that was defended by sense-data theorists. Phenomenal colors always appear as part of a visual field. A manifest aspect of the visual field is that it is spatial: when I experience a red patch, I see it with a certain shape. According to adverbialism, I sense (red and square)-ly. Now, it is highly plausible that only moving subjects can sense squarely or roundly. If this were true, a bodiless machine could not in principle produce its own visual field. Hence, the machine's inability to see spatially would entail its inability to see colorfully. If we accept this, we would be

committed to saying that a phenomenal experience of color is not something that happens in the head.

The view that the interaction between the body and its environment must be thought as essential parts of experiential processes is controversial. "*If there is one thing that scientists are reasonably sure of, it is that brain activity is both necessary and sufficient for biological sentience*", according to the neuroscientist Christof Koch (Koch 2004, p. 9). A good reason to believe that qualia are realized by neural processes is that there is long-standing evidence that conscious experiences can arise from the electrical stimulation of certain parts of the brain (Penfield 1958). The possibility of making a red spot appear in the visual field of a conscious subject by triggering some neurons strongly suggests that qualia are produced by the brain alone. Moreover, even without scientific proof, most of us know from dreaming experiences that qualia are independent of the interaction with the environment. Indeed, although dreams are arguably phenomenologically different from waking experiences, it is difficult to deny that there exist such things as dreamed qualia. The latter are generally not accessible to reflexive awareness but they are in *lucid* dreams where the dreamers are in full possession of their waking capacities (Gackenbach & LaBarge 2012). For experienced lucid dreamers, even the claim that the softness of a sponge is the result of an interaction with a real sponge may seem counterintuitive because they know that they could live this exact experience while lying still in bed with their eyes closed. Does it prove that having a body is not a necessary condition for qualia after all?

At this point, it is important to introduce a distinction between *causal* conditions and *constitutive* conditions for experience:

> *For example, cerebral blood flow is causally necessary for consciousness, but activation of the upper brainstem is much more plausibly a constitutive condition, part of what it is to be conscious.* (Block 2007)

Block's example is interesting because it reveals an important ambiguity in the context of the current discussion: how to determine what counts as "part of" an experiential process? Block suggests that the constitutive conditions of conscious experiences determine what they essentially are. To make an analogy, the radiation of the sun is causally necessary for the existence of a sunburn, but the sunburn itself is wholly constituted by properties of the skin.

Hence the sun is not part of what a sunburn essentially is. In the same way, although a real sponge is a causal condition for an experience of sponge softness, the softness quale itself may be wholly constituted by brain states. This would explain why, by reproducing the relevant causal inputs to the brains with electrodes, it would in principle be possible to induce a hallucinatory experience of softness.

The distinction between the constitutive and the causal basis of qualia echoes the distinction that I made in the first part between the correspondence problem and the problem of artificial qualia. Consciousness scientists are primarily interested in the constitutive basis of qualia in the sense that they try to identify what is essentially different between a human subject having qualitative experiences and one that does not. It turns out that the difference is expressible in terms of neural activity. However, it cannot be inferred from that observation that qualia are wholly constituted by brain states. It is a mistake to compare qualia to a sunburn. What can be compared to a sunburn in this context is the vehicle of the representation, a neural activation in the brain. When I squeeze a sponge and experience its softness, there is an internal state in my brain that represents the sponge as being soft. Does it mean that the quale of softness is a property of my brain? No, because a quale of softness is not a property, it is a holistic token process as argued in part 2.4. It is not possible to abstract the softness of the experience from the field in which it appears. This would mean that a quale of softness is constituted by a very complex causal process that starts with the squeeze of the sponge and ends in the brain. On this view, what really "produces" a quale could only be determined by the whole causal chain that resulted in this particular token experience.

Now, the problem of looking for the causal substrate of a quale is that it leads to a slippery slope, which in a sense goes in the opposite direction to the one discussed in the previous part about functionalism. To understand why, let's get back to the previous example of dreams. It is highly likely that dreamed qualia are causally dependent on real life experiences. Someone that has never felt the softness of a sponge in real life will hardly be able to dream this particular qualitative experience. This entails that the causal substrate of qualia can include events that are very far in the past. Indeed, I am now committed to the view that my dreamed quale of squeezing a sponge tonight is realized by a process that started a year ago when I squeezed a real sponge. This is for is claimed by Manzotti, who defends an externalist view of perception close to that of the sensorimotor approach:

> *Last night, I dreamed of my grandmother, who died in 1985. During my dream I had qualia connected to her. How was this possible? A possible answer is that any phenomenal experience is continuous with a physical event. Any phenomenal experience is a process that ends in a brain. How long could this process last? During normal perception, it seems acceptable to take into account a process that spans time and space. Visual perception requires a time span ranging from approximately 10 ms to 200 ms. Is there any scientific evidence that constrains the maximum time length of a process? As far as I know, there is none. Therefore, I suggest that dreams are just cases of postponed perception. My dream of my grandmother would be nothing else than a perception of my grandmother that took many years instead of a few milliseconds to be completed.* (Manzotti 2008)

Moreover, what is true for this temporal extension of the causal basis of qualia is also true for their spatial extension. When one looks at the stars, it could be claimed that the visual experience is a causal process that includes two objects separated by several light years. The problem with causality in general is that it provides no clear boundary separating what is part of the experiential process and what is not.

So why not just accept this absence of boundaries? Why not forget the idea that there is a clear separation between experience and non-experience? When I see a cat looking at an apple, I am looking at an experiential process. The process includes the eyes of the cats and its brain, but it also includes the apple and the light waves that are reflected by the surface of the apple. A simple way to see this is to remark that interfering with the real apple modifies the experience of the cat. I don't need to manipulate the brain of the cat to change the phenomenal content of its experience, I just need to put something in its visual field. And now that the frontier between experiential processes and non-experiential processes is abolished, the problem of artificial experience becomes intractable because there is no way to individualize experience. An experiencing mind cannot be precisely localized, there is no way we can point to a process and say "*this is the experience*". To understand why, Alva Noë proposes an analogy with money:

> *As a comparison, consider that there's nothing about this piece of paper in my hand, taken in isolation, that makes it one dollar. It would be ludicrous to search for the physical or molecular correlates of its monetary value. The monetary value, after all,*

> *is not intrinsic to the piece of paper itself, but depends on the existence of practices and conventions and institutions. The marks or francs or pesos or lire in your wallet didn't change physically when, from one day to the next, they ceased to be legal tender. The change was as real as it gets, but it wasn't a physical change in the money. Maybe consciousness is like money. Here's a possibility: my consciousness now—with all its particular quality for me now— depends not only on what is happening in my brain but also on my history and my current position in and interaction with the wider world.* (Noë 2009, p. 4)

This is an attractive view because if externalism of this sort is true, it would explain why the problem of qualia is so hard. We will never model experience itself because modeling experience would boil down to modeling the external world itself. Indeed, if experience is an access to external reality, it is not surprising that the structure of experience reflects the structure of the external world that is accessed. We will not find psychophysical laws relating qualia to external processes of the world because the two emerge together. As Varela, Thompson and Rosch puts it:

> *We propose as a name the term* enactive *to emphasize the growing conviction that cognition is not the representation of a pregiven world by a pregiven mind but is rather the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs. The enactive approach takes seriously, then, the philosophical critique of the idea that the mind is a mirror of nature but goes further by addressing this issue from within the heartland of science.* (Varela & al. 1993)

In the next part, I will discuss the *enactive* or *embodied* approach. According to the defenders of this view, only a *living* creature can experience the world. The fact that we cannot clearly locate consciousness does not entail that it has always existed. Indeed, even if we cannot precisely locate an experiential process, it does not mean that there always has been experience. Presumably, qualia appeared in the history of evolution.

3.4. Life and the evolutionary origins of qualia

There are good reasons to believe that qualia have evolutionary origins. The unpleasant way of accessing body damage is not unpleasant by mere coincidence. Its qualitative aspect is related to the survival advantages that it provides, as noted by William James:

> *It is a well-known fact that pleasures are generally associated with beneficial, pains with detrimental, experiences. All the fundamental vital processes illustrate this law. Starvation, suffocation, privation of food, drink and sleep, work when exhausted, burns, wounds, inflammation, the effects of poison, are as disagreeable as filling the hungry stomach, enjoying rest and sleep after fatigue, exercise after rest, and a sound skin and unbroken bones at all times, are pleasant. Mr. Spencer and others have suggested that these coincidences are due, not to any pre-established harmony, but to the mere action of natural selection which would certainly kill off in the long-run any breed of creatures to whom the fundamentally noxious experience seemed enjoyable.* (James 1890, pp. 143-144)

Qualia as we know them today are biological phenomena. Since "*nothing in biology makes sense except in the light of evolution*" (Dobzhansky 1973), it is relevant to examine qualia from the point of view of evolutionary biology.

Thinking about the way natural selection shaped conscious experiences is interesting in several ways. To understand why, let's start by taking vision as an example. The way humans see is complex but cognitive science is able to explain a lot about vision in computational terms (Marr 1982). Not only do we have good models of vision but we are even able to build artifacts that visually perceive their environment. Computer vision as it exists today is perfectly functional, arguably even more performant than human vision for certain tasks (Yu & al. 2017). The problem is that, as far as we know, artificial vision is not phenomenal. A computer, a robot or a self-driving car do not have their own visual field full of colors and shapes. So now we are in an odd situation because it is as if we solved the problem of vision but somehow the solution that we arrived at is different from the one performed by nature in at least one important regard. We wonder what it is that we have to add to a perceiving machine so that it begins to have visual experiences. The issue is that asking the question in those terms makes it manifest that the experiential aspect that we may add would be useless.

The machine is already able to detect colors and shapes, so why would it need to experience them? This is exactly how we arrive at the conclusion that qualia are mysterious *epiphenomenal* properties (Robinson 2019) or that they do not really exist.

Instead of wondering what sort of information processing could be added to a computer vision program so that it starts to have color experiences, a better idea might be to think about the way conscious vision emerged in nature. If we go up the phylogenetic tree, we might expect that our common ancestors with other animals had more rudimentary visual experiences. If we continue to climb the tree, we may arrive at creatures that have no color or shape experience. However, these creatures may have non-visual experiences. Considering the genesis of visual experience allows us to picture how our experiences may have been progressively built up by nature in the same way that our bodies were. The point is that nature may have solved the problem of vision from a basis that was already phenomenal. If this were true, it would probably mean that we would never account for the qualitative content of experience from the complex information processes happening in the parts of the brain that appeared last. Instead of conceiving the latter as the generator of qualia, it is not unreasonable to consider that their role is rather to modulate pre-existing phenomenal experience. Hence it would be no surprise that the replication of the high-level cortical processes in computer vision systems do not involve visual qualia. Visual experience may be hierarchically dependent on more fundamental qualia, as argued in part 2.5. These more fundamental qualia might be instances of "raw" feelings:

> *The advent of feelings was simultaneously the advent of the mind. Early organisms capable of feeling were, for the first time in evolution and unlike all other life forms, aware of some aspects of their own existence. Feelings paved the way for the establishment of higher levels of cognition and consciousness, culminating in the modern human mind. Accordingly, shedding light on the underpinnings of feeling is likely to provide insights into consciousness and the mind.* (Damasio & Carvalho 2013)

According to the neuropsychologist Antonio Damasio, cognitive science focused too much on the complex thinking abilities of the human mind and neglected its affective dimension, what is sometimes called *affective consciousness* (Panksepp 2007). Damasio argues that this

mistake has very old roots in Western philosophy, as illustrated by Descartes' "*I think therefore I am*":

> *Taken literally, the statement* [of Descartes] *illustrates precisely the opposite of what I believe to be true about the origins of mind and about the relation between mind and body. It suggests that thinking, and awareness of thinking, are the real substrates of being. And since we know that Descartes imagined thinking as an activity quite separate from the body, it does celebrate the separation of mind, the "thinking thing" (res cogitans), from the non thinking body, that which has extension and mechanical parts (res extensa).* (Damasio 1994, p. 248)

Damasio claims that the activity of the body should be thought as part of the constitutive basis of an experiential process in virtue of being a *living* body. Accordingly, the key to solving the problem of qualia lies in the specific properties of the process that we call life. On this view, a robot cannot experience the world because it is not alive.

According to Varela and Maturana, the crucial difference between computational and living processes is that the latter are *self-generative* or *autopoietic*:

> *an autopoietic machine continuously generates and specifies its own organization through its operation as a system of production of its own components* [...]
> *Therefore an autopoietic machine is an homeostatic system which has its own organization (defining network of relations) as the fundamental variable which it maintains constant.* (Varela & Maturana 1980, p. 79)

Their claim is that there is an important difference between the causal processes realized by computational machines and the ones realized by organisms. This is supposedly because the latter are "*agents*" in the sense that they determine the way they behave in response to stimuli. "*An agent is a source of activity, not merely a passive sufferer of the effects of external forces*" (Barandarian & al. 2009). While a programmed robot always reacts to its environment by mere "reflex", a living agent would be able to enact complex and circular causal processes:

> *a successful action has a 'circular' form; it begins and ends in the same entity within the agent. In contrast, a reflex is a 'linear', programmed movement that, once initiated, is carried out irrespective of its effect (if any) on that which triggered it. Unlike a reflex, an action has a homeostatic nature; an agent moves to keep itself in a certain state. And, as argued above, it is a substance that determines its own movement. Hence, an agent is a material having a homeostatic nature.* (Longinotti 2017)

What is life? From the perspective of physics, living systems are "*no more than a manifestation of a set of complex chemical reactions and, as such, are governed by the rules of kinetics and thermodynamics*" (Pross 2003). However, one of the notable characteristics of living systems is their unique capacity to resist entropy, that is to maintain a certain level of internal organization without external intervention (Schrödinger 1944). This self-maintained organized state is called *homeostasis.* A homeostatic system, such as a living cell, continually maintains its integrity through *metabolic reactions*, the biochemical processes permitting energetic exchange with the external environment. According to some advocates of the embodied approach to cognition, it is in virtue of being constantly in action in order to stay itself that a living system is able to feel. From this perspective, affective qualia such as pain and pleasure are *signals* that inform the subject of "*deviations from homeostatic set-points*" (Solms 2014). This is an interesting position because it links the causal powers of qualia to their qualitative aspect. Qualia are informational processes, as argued in part 1.2: there is phenomenal information in the sense that it is the qualitative experience itself that informs the subject. Moreover, the difference between a robot and a living organism is that the latter is "self-individuated". This could explain why an organism has its own subjective qualia. A minimal form of self could be explained by agency and homeostatic processes: an organism continuously maintains its boundary with the external environment by a process of self-generation. By contrast, a computational process is only individuated from an external perspective, it does not have its own point of view.

The problem with the embodied approach is that it is not clear at all how and why the metabolic reactions underlying homeostasis are responsible for feelings. The view that there is something special with biochemistry with regard to experience seems like an "*implausible vitalistic claim*", as Dennett puts it. After all, even though living organisms are arguably complex systems manifesting circular causal processes, they are still complex organizations

of matter. The main difference with functionalism as it is traditionally articulated is that the necessary organizational and functional properties required for qualia to exist should be specified at a finer-grained level, namely that of biochemistry. Seen in this light, the embodied approach is close to Searle's biological naturalism, the view that "*the right level to account for the very existence of consciousness is the biological level*" (Searle 2007). Although shifting the problem of qualia to biology may be a legitimate move, the main limitation of functionalism still applies. Indeed, biology provides functional explanations and it is hard to see how a functional explanation can in principle account for experience. This was in fact the whole point of Chalmers' hard problem, who argued that "*there is no cognitive function such that we can say in advance that explanation of that function will automatically explain experience"* (Chalmers 1995a).

However, the point of the advocates of the embodied approach is that what applies for *cognitive* functions may not be true for *feelings,* as argued by the neuropsychologist Mark Solms:

> *Would he* [Chalmers] *have said such things in the first place if he had been talking about affective – rather than cognitive – functions? [...] I can say in advance that explanation of the function of feeling will automatically explain experience.* (Solms 2021, pp. 396-397)

When seen from biology instead of cognitive science, Chalmers' Hard problem boils down to the question of explaining why feelings are felt. Expressed in this way, the problem seems absurd. It is in the intrinsic nature of a feeling to be felt qualitatively, so if we understand the function that feelings fulfill, then we have explained everything there is to explain about feelings. Is this really a satisfying answer? I believe it is just a way of begging the question. The main reason why the Hard problem is hard is that it is very obscure how inert matter can possibly produce feelings with an intrinsic "what-it's-like-ness". The move of advocates of the embodied approach such as Damasio and Solms is to point out that the "what-it's-like-ness" of feelings is related to its intrinsic valence: pain is intrinsically bad and pleasure is intrinsically good. Admittedly, it is reasonable to think that these valenced phenomenal states appeared for evolutionary reasons because it is indeed useful for an organism to immediately feel what is good or bad for its fitness. However, this hardly explains how matter became feeling. The traditional mind-body problem is always the same,

it is still unintelligible how we go from material processes to valenced qualitative states such as pain.

The mind-body problem can be decomposed into two parts that Evan Thompson calls the "*mind-mind problem*" and the "*body-body problem*" (Thompson 2010, pp. 6-7). The mind-mind problem, as coined by Ray Jackendoff (1987, p. 20), concerns the relation between the computational mind and the phenomenological mind. Thanks to cognitive scientists, we have a relatively good understanding of the subpersonal computational processes happening in the brain, and we are able to partially reproduce them artificially. The phenomenological mind refers to conscious experience, qualia, what-it's-like-ness. It is not separated from the computational mind in humans. When humans see, the process is both computational and phenomenological but we don't fully understand how the two relate to each other. To put it simply, we need to understand how to go from "raw" affective states to more complex structured experiences. This is arguably a huge scientific challenge but it is not impossible to solve. It is reasonable to suppose that neuroscientific theories of consciousness[9] will greatly help to solve the mind-mind problem. There is no doubt that brain processes constitute the most important part of what an experience is and thus a full comprehension of the brain will allow us to understand how human consciousness works. However, even if the mind-mind problem is very difficult, it is still an "easy" problem. The real Hard problem of consciousness is the body-body problem that consists in explaining how to go from an *organic* body to a subjectively *felt* body:

> *In this formulation of the hard problem, I have substituted the term body for physical. Body connotes life, a living organism, and is richer in meaning than physical in the Cartesian sense. Drawing on this richness can help us to refine the terms of the explanatory gap.* (Thompson 2010, p. 235)

What Thompson means by "*physical in the Cartesian sense*" is in fact "*mechanistic*". In the philosophy of biology, there is a long-standing opposition between organicists and mechanists that has roots in the nineteenth century debate around vitalism (Allen 2005). There may be something holistic in living processes that could possibly help to explain why only an organic brain connected to a body can produce the unified experiential "field".

[9] Such as the global workspace theory of consciousness (Dehaene & al. 2014) or the integrated information theory (Tononi & al. 2016)

However, even when accepting some form of anti-reductionist organicism according to which biological processes are very special kinds of non-linear causal chains, it is still unintelligible how qualia emerged in a universe completely devoid of experience.

3.5. Panpsychism

As already discussed in part 1.7, physicalism is the view that everything that exists in the universe is physical. Thus, according to physicalists, either qualia are physical processes or they do not exist. Most physicalists want to keep intact the pre-theoretical intuition that phenomenal experiences are real. Indeed, few of them are willing to explicitly deny the phenomenal character of pain. The problem is that, whatever this phenomenal character really is, it does not seem to be of the same nature as physical processes. Hence physicalists face what is famously called *the explanatory gap* (Levine 1983). On one side of the gap there are the experiential processes such as the taste of coffee or the redness of red and on the other side there are the physical processes that are described by the natural sciences such as gravitation, metabolic reactions and cortical activity. If the latter are "physical", it is intuitively hard to figure out how the former could be. Joseph Levine remarks that "*this kind of intuition about our qualitative experience seems surprisingly resistant to philosophical attempts to eliminate it*" (ibid.). Why is it intuitively difficult to accept that qualia are physical? In a pre-theoretical and intuitive sense, the word "physical" is usually defined *negatively*, precisely by contrast with "mental". For instance, in the Oxford Dictionary it is defined as "*relating to the body as opposed to the mind*" (Physical, n.d.). Physical things are *by definition* those that do not depend on any subjective experience, thus experience itself obviously cannot be physical in this sense.

There is a more interesting way to understand physicalism. According to Daniel Stoljar, one way to interpret physicalism is to claim that it is true "*if and only if every instantiated fundamental property is physical*" (Stoljar 2010, p. 35). Stoljar invites us to think of fundamental properties as "*the ingredients of the world*" (p. 34), then uses a theological metaphor to illustrates what he means:

> *Imagine God as a divine cake maker: to make the cake of the world what he would have to do would be to arrange the basic ingredients in the right way, and everything else would follow immediately.* (ibid. , pp. 34-35)

From this perspective, the essential claim of physicalists is that experience is not "*fundamental*". The explanatory gap is thus the problem of explaining experience with "*basic ingredients*" that are in and of themselves non-experiential. At first glance, it might not seem to be such an insurmountable problem. There are indeed natural phenomena that are scientifically explainable in terms of basic elements that do not individually exhibit the properties of the explanandum. Interestingly, an example that is often put forward by physicalists in this context is that of *life*. Back in the nineteenth century, vitalists believed that living organisms were "*fundamentally different*" from non-living entities because they were thought to "*contain some non-physical elements"* (Bechtel & Williamson 1998). Most scientists today believe that vitalism is false as the advances in molecular biology tend to show that there is nothing more to life than very complex physical mechanisms. In other words, life *emerged* from more fundamental physical phenomena. According to physicalists, qualia also emerged from physical processes. From this perspective, those who believe that there is a Hard problem of consciousness are the new vitalists, they think that we need an "*extra-ingredient*" (Chalmers 1995a) to fully explain experiential processes, but the future will probably prove them wrong.

The problem is that it is not obvious at all that qualia can be understood by analogy with life. Galen Strawson proposes to think of experience as being more akin to space (Strawson 2006). He argues that it would be inconceivable that spatial entities emerged from more fundamental non-spatial entities. It would appear miraculous that when wholly non-spatial phenomena entertain certain non-spatial relations, they give rise to spatial phenomena. This latter example would be a case of "*brute emergence*":

> *Emergence can't be brute. It is built into the heart of the notion of emergence that emergence cannot be brute in the sense of there being absolutely no reason in the nature of things why the emerging thing is as it is (so that it is unintelligible even to God). For any feature Y of anything that is correctly considered to be emergent from X, there must be something about X and X alone in virtue of which Y emerges, and which is sufficient for Y.* (ibid.)

Strawson's argument mainly relies on the premise that, as for space, there is something "fundamental" with experiential phenomena. His conclusion is that qualia must be part of the

"*ingredients of the world*", a view called *panpsychism*. It could be argued that Strawson is just begging the question. Indeed, he concludes that qualia cannot be emergent on the premise that they are fundamental. But in this case, physicalists are also begging the question when insisting that, whatever experience is, it is not "*fundamental*". After all, we have no convincing theory of qualia today, so why not suppose that our future explanation of consciousness will consider it fundamental? As Noam Chomsky argues, what is considered "*fundamental*" by science changes over time:

> *Repeatedly, the more "fundamental" science has had to be revised, sometimes radically, for unification to proceed. Suppose that a nineteenth century philosopher had insisted that "chemical accounts of molecules, interactions, properties of elements, states of matter, etc. must in the end be continuous with, and harmonious with, the natural sciences," meaning physics as then understood. They were not, because the physics of the day was inadequate. By the 1930s, physics had radically changed, and the accounts (themselves modified) were "continuous" and "harmonious" with the new quantum physics.* (Chomsky 2000, p. 82)

Chomsky's point is related to Hempel's argument against physicalism. Hempel (1969) proposed a dilemma to physicalists: either they define fundamentality with respect to our current conception of physics, or they mean it in the sense of what a future ideal physics will provide as the ingredients of the world. In the former case, physicalism is most likely false since our contemporary physics is not complete and in the latter case physicalism is tautological because an ideal physics will explain everything by hypothesis. So can we seriously consider the possibility that future scientists will include qualia as part of the fundamental ingredients of the world?

There are different ways of understanding that experience is "fundamental". Strawson (2006) proposes a form of "*micropsychism*" or "*microidealism*":

> *Micro-idealism is the thesis that all concrete facts are grounded in facts about the mental states of (or mentality associated with) fundamental microscopic entities, such as quarks or photons.* (Chalmers 2019).

There are several problems with this view. First, it is probably not compatible with our current understanding of physics. Indeed, fundamental particles in contemporary Quantum Field Theory are not understood as individual objects that could be the bearers of intrinsic mental "properties" (Ryder 1996). Moreover, "*even in a classical physics framework, there are challenges, the first among which is the challenge of space and time*" (Chalmers 2019). I put aside these objections since there may be ways to answer them. A more important obstacle to micropsychism is called the *combination problem*: how do micro-experiences combine to form a macro-experience? (Seager 1995) It is a metaphysical puzzle that may be almost as hard as the Hard problem itself. Moreover, it is not compatible with the adverbial analysis of experience that I proposed in this work. Indeed, I argued that the phenomenal content of experience should not be reified. An experience of a red square cannot be analyzed as an individual red sense-datum that is combined with an individual square sense-datum. Instead, there is an adverbial modification that can be analysed as an adverbial modification of a field of experience.

A panpsychist alternative to micropsychism is called "*cosmopsychism*" or "*cosmic idealism*":

> *Cosmic idealism is the thesis that all concrete facts are grounded in facts about the mental states of (or the mentality associated with) a single cosmic entity, such as the universe as a whole or perhaps a god.* (Chalmers 2019)

Different versions of cosmopsychism have been defended, for example by Bernardo Kastrup (2019) and Philip Goff (2019). In their view, there is no separation between experience and the physical world because the whole universe is one holistic mental process. There are several obstacles to this view, one of them being the difficulty to explain individual experiences. Idealists of this sort will typically say that we *believe* that we are separated experiential processes, but that in reality we are not. In this respect, this position has similarities with Hindu and Buddhist idealism and it undeniably has a mystical dimension. A full examination of these sorts of metaphysical positions and of all the other variants of panpsychism would go beyond the scope of this work.

It is not surprising that taking the Hard problem seriously inevitably leads to radical metaphysical positions such as panpsychism. If one believes that some material processes of the outside world are qualia, it becomes impossible to clearly distinguish between what is

mental and what is not without introducing a mysterious gap between the two. Now, if one version of panpsychism is true and that all processes of the world are qualia, there is a sense in which we find ourselves in the same situation as that of elminiativists and illusionists: there is no fundamental difference between us and machines because there is no real separation between experiential and non-experiential processes.

Conclusion

The main conclusion that can be drawn from this work is that we are very far from having a full understanding of qualia. As argued in the first part, the problem of machine consciousness is essentially a philosophical problem about the metaphysical status of qualitative experiences. Neuroscience has undeniably contributed to a huge progress in our understanding of the role of the brain in conscious cognition. However, asking what it would take to build a robot that feels pain quickly leads to complicated questions about the nature of qualia. In the second part, I argued that the phenomenal content of experience cannot easily be analyzed as individual objects or properties. Objects and properties are represented in experience and they are represented in a certain way. We can become aware of these modes of presentation but it is very difficult to understand their ontological status. In the third part, I argued that there are good reasons to believe that the content of experience is probably not something that is completely happening inside the head. An experience of the world is a process in which the world itself is included, and there is no clear separation between what belongs to an experience and what does not. Nevertheless, it is reasonable to suppose that qualia are processes that appeared at some point in the history of the universe. If this is the case, they probably arose for evolutionary reasons: it was useful for some of our distant ancestors to have the ability to feel. However, it is difficult to understand how biological phenomena happening in a word devoid of experience miraculously became feelings. This view seems to commit to a form of vitalism that is unacceptable for most physicalists. If we want to keep intact the idea that there is no ontological discontinuity in nature and avoid the explanatory gap, we must abolish the clear distinction between the mental and the physical. And there are not so many ways to do that: either we choose to eliminate qualia or we accept that they are fundamental. Both positions are radical and both lead to a dissolution of the problem of artificial qualia.

References

Aaronson, S. (2016). “Why I Am Not an Integrated Information Theorist (or, The Unconscious Expander).” The Blog of Scott Aaronson.

Allen, G. E. (2005). Mechanism, vitalism and organicism in late nineteenth and twentieth-century biology: the importance of historical context. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36(2), 261-283.

Armstrong, D. M. (1968). A materialist theory of the mind. Routledge.

Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision substitution by tactile image projection. Nature, 221(5184), 963-964.

Bach-y-Rita, P., & Hughes, B. (1985). Tactile vision substitution: some instrumentation and perceptual considerations. In Electronic spatial sensing for the blind (pp. 171-186). Springer, Dordrecht.

Bach-y-Rita, P. (1997). Substitution sensorielle et qualia. Perception et intermodalité. Approches actuelles de la question de Molyneux, 81-100.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. Cognition, 20(3), 191-208.

Balog, K. (2009). Phenomenal concepts.

Balog, K. (2012). Acquaintance and the mind-body problem. New perspectives on type identity: The mental and the physical, 16-42.

Barandarian, X., Di Paolo, E., Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. Adapt. Behav. 17(5), 370

Bayne, T. (2010). The unity of consciousness. Oxford University Press.

Bechtel, W., & Williamson, R. C. (1998). "Vitalism". In E. Craig (ed.). Routledge Encyclopedia of Philosophy. Routledge.

Bickle, J. (2020). "Multiple Realizability", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.)

Block, N. (1978). Troubles with functionalism.

Block, N. (1995). On a confusion about a function of consciousness. Behavioral and brain sciences, 18(2), 227-247.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. Behavioral and brain sciences, 30(5-6), 481.

Bourget, D. & Mendelovici, A. (2019). "Phenomenal Intentionality", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.)

Brentano, F. (1874). Psychology from an Empirical Standpoint (London: Routledge, 1995)

Brook, A. & Raymont P. (2017). "The Unity of Consciousness", The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Edward N. Zalta (ed.)

Carnap, R. (1928). The logical structure of the world. Translated by George, R. A. (1967) London: Routledge.

Carnap, R. (1936). Testability and meaning. Philosophy of science, 3(4), 419-471.

Chalmers, D. J. (1995a). Facing up to the problem of consciousness. Journal of consciousness studies, 2(3), 200-219.

Chalmers, D. J. (1995b). Absent Qualia, Fading Qualia, Dancing Qualia. Conscious experience, 309.

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford Paperbacks.

Chalmers, D. J. (2000). What is a neural correlate of consciousness. Neural correlates of consciousness: Empirical and conceptual questions, 17-39.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. Journal of Cognitive Science, 12(4), 325-359.

Chalmers, D. J. (2018). The meta-problem of consciousness.

Chalmers, D. (2019). Idealism and the mind-body problem.

Chisholm, R. (1957). Perceiving: A Philosophical Study. Ithaca: Cornell University Press.

Chomsky, N. (2000). New horizons in the study of language and mind. Cambridge University Press.

Chudnoff, E. (2013). Gurwitsch's phenomenal holism. Phenomenology and the Cognitive Sciences, 12(3), 559-578.

Churchland, P. (2007). On the reality (and diversity) of objective colors: How color-qualia space is a map of reflectance-profile space. Philosophy of Science, 74(2), 119-149.

Clark, A. (2000). A theory of sentience. Clarendon press.

Clark, A., & Chalmers, D. (1998). The extended mind. analysis, 58(1), 7-19.

Crane, T. (1998). Intentionality as the mark of the mental.

Crane, T. (2000). The origins of qualia.

Crick, F. and Koch, C. (1990). Towards a neurobiological theory of consciousness. Seminars in the Neurosciences 2: 263-275.

Crick, F. & Koch, C. (1998). Consciousness and neuroscience. Cerebral Cortex 375: 121-123

Cytowic, R. E. (1989). Synesthesia and mapping of subjective sensory dimensions. *Neurology*, *39*(6), 849-850.

Dainton, B. (2010). Phenomenal holism. Royal Institute of Philosophy Supplement, 85(67), 113.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain.

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: evolutionary and neurobiological origins. Nature reviews neuroscience, 14(2), 143-152.

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. Curr. Opin. Neurobiol. 25, 76–84.

De Mol, L. (2019). "Turing Machines", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.)

Dennett, D. C. (1978). Why you can't make a computer that feels pain. Synthese, 38(3), 415-456.

Dennett, D. C. (1988). Quining qualia. In Consciousness in modern science. Oxford University Press.

Dennett, D. C. (2004). "Epiphenomenal" Qualia. There's something about Mary: essays on phenomenal consciousness and Frank Jackson's knowledge argument, 59-73.

Dennett, D. C. (2006). What robomary knows.

Dinets, V. (2016). No cortex, no cry. Animal Sentience, 13(Commentary on Key on Fish Pain).

Dobzhansky, T. (1973). "Nothing in Biology Makes Sense Except in the Light of Evolution", American Biology Teacher, 35 (3): 125–129

Dretske, F. (1997). Naturalizing the mind. MIT Press.

Dretske, F. (2003). Experience as representation. Philosophical issues, 13, 67-82.

Ducasse, C. J. (1942). Moore's refutation of idealism.

Edelman, G. M., & Tononi, G. (2013). Consciousness: How matter becomes imagination. Penguin UK.

Feynman, R. (1988). Richard Feynman's blackboard at time of his death. https://digital.archives.caltech.edu/islandora/object/image%3A2545

Flor, H. (2002). Phantom-limb pain: characteristics, causes, and treatment. The Lancet Neurology, 1(3), 182-189.

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

Frege, G. (1948). Sense and reference. The philosophical review, 57(3), 209-230.

Freud, S. (1915). The Unconscious in Psychoanalysis, in Collected Papers, vol. 4. pp. 98-136. J. Riviere tr. New York: Basic Books, 1959.

Gackenbach, J., & LaBarge, S. (Eds.). (2012). Conscious mind, sleeping brain: Perspectives on lucid dreaming. Springer Science & Business Media.

Gennaro, R. (2004). Higher-Order Thoughts, Animal Consciousness, and Misrepresentation: A reply to Carruthers and Levine. In RJ Gennaro (Ed.) Higher-Order Theories of Consciousness. Amsterdam and Philadelphia: John Benjamins.

Gert, J. (2020). Information-Theoretic Adverbialism. Australasian Journal of Philosophy, 1-20.

Goff, P. (2019). Cosmopsychism, Micropsychism and the Grounding Relation. In The Routledge Handbook of Panpsychism (pp. 144-156). Routledge.

Goldstein, I. (1989). Pleasure and pain: Unconditional, intrinsic values. Philosophy and Phenomenological Research, 50(2), 255-276.

Gregory, R. L., & Wallace, J. G. (1963). Recovery from early blindness. Experimental psychology society monograph, 2, 65-129.

Haeckel, E. (1892). Our Monism. The principles of a consistent, unitary world-view. The Monist, 481-486.

Haugeland, J. (1989). Artificial intelligence: The very idea. MIT press.

Hempel, C. (1969). Reduction: Ontological and Linguistic Facets, in S. Morgenbesser, et al. (eds.), Essays in Honor of Ernest Nagel, New York: St Martin's Press.

Hume, D. (1740). A treatise of human nature. Volume 1: Texts. Oxford University Press 2007

Husserl, E. (1931). Méditations cartésiennes: introduction à la phénoménologie.

Husserl, E. (1958). Die Idee der Phänomenologie. Husserliana II. W. Biemel (Ed.). (2nd ed.) Martinus Nijhoff: The Hague. English translation: The idea of phenomenology. Edmund Husserl: Collected Works, Vol VIII. (L. Hardy, Trans.).

Jackendoff, R. (1987). Consciousness and the computational mind. The MIT Press.

Jackson, F. (1975). On the adverbial analysis of visual experience. Metaphilosophy, 6(2), 127-135.

Jackson, F. (1982). Epiphenomenal qualia. The Philosophical Quarterly (1950-), 32(127), 127-136.

Jackson, F. (1986). What Mary didn't know. The Journal of Philosophy, 83(5), 291-295.

Jackson, F. (2004). There's something about Mary: essays on phenomenal consciousness and Frank Jackson's knowledge argument. MIT press

Jackson, F., Pargetter, R. and Prior, E. (1982). Functionalism and Type-Type Identity Theories, Philosophical Studies, 42: 209–225.

Jacob, P. (2019). "Intentionality", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.)

James, W. (1890). The principles of psychology (Vol. 1). New York: Henry Holt.

Kammerer, F.. (2016). “The hardest aspect of the illusion problem — And how to solve it.” 23. 124-139.

Kammerer, F. (2019). Conscience et matière. Une solution matérialiste au problème de l'expérience consciente. Editions Matériologiques.

Kandinsky, W. (1911). Concerning the spiritual in art. Translated by Michael T. H. Sadler (2008). The Floating Press.

Kastrup, B. (2019). Analytic Idealism: A consciousness-only ontology.

Kobes, B. W. (1995). Access and what it is like. Behavioral and Brain Sciences, 18(2), 260-260.

Koch, Christof (2004). The quest for consciousness: a neurobiological approach. Englewood, US- CO: Roberts & Company Publishers.

Köhler, W. (1967). Gestalt psychology. Psychologische Forschung, 31(1), XVIII-XXX.

Kuehni, R. G. (2003). Color Space and its Divisions, New York: Wiley.

Kurzweil, R. (2002). Locked in his Chinese Room, in Richards 2002, 128–171.

La Rochefoucauld, F. (1678). Maximes et réflexions morales.

Lau, H. and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences, 15(8):365–373.

Levin, J. (2018). "Functionalism", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.)

Levine, J. (1983). Materialism and qualia: The explanatory gap. Pacific philosophical quarterly, 64(4), 354-361.

Lewis, D. (1988). What experience teaches.

Linton, S. J. (2005). Understanding pain for better clinical practice: a psychological perspective.

Loar, B. (1997). Phenomenal states (Revised version). In Block N., Flanagan O. & Güzeldere G. (eds), The Nature of Consciousness, MIT Press.

Loar, B. (2004). Phenomenal States (Revised Version). There's Something about Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument, 219.

Locke, J. (1689). Essay Concerning Human Understanding. Oxford: Oxford University Press.

Longinotti, D. (2017). Agency, qualia and life: connecting mind and body biologically. In 3rd Conference on" Philosophy and Theory of Artificial Intelligence (pp. 43-56). Springer, Cham.

Lycan, W. G. (2019). Representational Theories of Consciousness, The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.)

Mandik, P. (1999). Qualia, space, and control. Philosophical Psychology, 12(1), 47-60.

Manzotti, R. (2008). A process-oriented view of qualia. The Case for Qualia, ed. by E. Wright, MIT Press, Cambridge, 175-190.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. AI magazine, 27(4), 12-12.

Merikle, P. (2000). Subliminal perception.

Merleau-Ponty, M. (1945), Phénoménologie de la perception. Gallimard

Metzinger, T. (2004). Being no one: The self-model theory of subjectivity. MIT Press.

Michel, M. (2019). Fish and microchips: on fish pain and multiple realization. Philosophical Studies, 176(9), 2411-2428.

Morgan, M. J. (1977). Molyneux's question: Vision, touch and the philosophy of perception. Cambridge U Press.

Noë, A. (2009). Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness. Macmillan.

O'Regan, J. K. (2011). Why red doesn't sound like a bell: Understanding the feel of consciousness. Oxford University Press.

Panksepp, J. (2007). Affective consciousness. The Blackwell companion to consciousness, 114-129.

Peirce, C. S. (1866). "Lowell lecture, ix." Writings of Charles S. Peirce: A chronological edition. M. H. Fisch. Ed. Bloomington, Indiana, Indiana University Press. I, 1857-1866: 471-86.

Penfield, W. (1958). Some mechanisms of consciousness discovered during electrical stimulation of the brain. Proceedings of the National Academy of Sciences of the United States of America, 44(2), 51.

Piaget, J. (1937). The construction of reality in the child. Translated by Cook M. (2013). Routledge.

Pitt, D. (2020). Mental Representation, The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.)

Pross, A. (2003). The driving force for life's emergence: kinetic and thermodynamic considerations. Journal of theoretical Biology, 220(3), 393-406.

Putnam, H. (1967). Psychological predicates. Art, mind, and religion, 1, 37-48.

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., ... & Vader, K. (2020). The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. Pain, 161(9), 1976-1982.

Rescorla, M. (2020). "The Computational Theory of Mind", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.)

Robinson, W. (2019). "Epiphenomenalism", The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), Edward N. Zalta (ed.)

Rochat, P. (2011). What is it like to be a newborn?. The Oxford handbook of the self (Chapter 2). Gallagher, S. (Ed.). Oxford University Press.

Russell, B. (1912). The problems of philosophy. Oxford University Press, 2001.

Ryder, L. H. (1996). Quantum field theory. Cambridge university press.

Sacks, O. & Wasserman, R. (1987). “The case of the colorblind painter”. New York Review of Books, November 19, 25-34.

Schrodinger, E. (1944). What is life? The physical aspect of the living cell.

Schwitzgebel, E. (2015). If materialism is true, the United States is probably conscious. Philosophical Studies, 172(7), 1697-1721.

Seager, W. (1995). Consciousness, information and panpsychism. Journal of Consciousness Studies, 2(3), 272-288.

Searle, J. R. (1980). Minds, brains, and programs. & &, 417-457.

Searle, J. (1992), The Rediscovery of the Mind. Cambridge, MA: The MIT Press.

Searle, J. R. (2004). Comments on Noë and Thompson, ‘are there neural correlates of consciousness?’ J Consciousness Stud 11(1):80-82

Searle, J. R. (2005). Consciousness: What we still don’t know. The New York Review of Books, 52(1), 36-39.

Searle, J. R. (2007). Biological naturalism. The Blackwell companion to consciousness, 325-334.

Searle, J. R. (2018). The Philosophy of Perception and the Bad Argument. E. Felder, & A. Gardt, Wirklichkeit oder Konstruktion, 66-76.

Searle, J. R., Dennett, D. C., & Chalmers, D. J. (1997). The mystery of consciousness. New York Review of Books.

Seibt, J. (2020). "Process Philosophy", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.)

Sellars, W. (1975). The adverbial theory of the objects of sensation. Metaphilosophy, 6(2), 144-160.

Shields, C. (2011). On behalf of cognitive qualia. Cognitive Phenomenology, 215-235.

Shoemaker, S. (1975). Phenomenal similarity. Crítica: Revista Hispanoamericana de Filosofía, 3-37.

Shoemaker, S. (1982). The Inverted Spectrum. Journal of Philosophy, 79: 357–81;

Sloman, A. and Logan, B. (1998). Architectures for human-like agents. Paper presented to European Conference on Cognitive Modelling, Nottingham, April.

Smart, J. J. (1959). Sensations and brain processes. The Philosophical Review, 68(2), 141-156.

Smart, J. J. C. (2017). "The Mind/Brain Identity Theory", The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.)

Solms, M. (2014). A neuropsychoanalytical approach to the hard problem of consciousness. Journal of integrative neuroscience, 13(02), 173-185.

Solms, M. (2021). The hidden spring: A journey to the source of consciousness. WW Norton & Company.

Stanley, R. P. (1999). Qualia space. Journal of Consciousness Studies, 6(1), 49-60.

Stoljar, D. (2010). Physicalism. Routledge.

Strawson, G. (1994). Mental reality. mit Press.

Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. Journal of consciousness studies, 13(10-11), 3-31.

Thompson, B. (2009). Senses for senses. Australasian Journal of Philosophy, 87(1), 99-117.

Thompson, E. (2010). Mind in life. Harvard University Press.

Tononi, G. (2004). "An Information Integration Theory of Consciousness". BMC Neuroscience 5 (1): 42.

Tononi, G. (2014). Why Scott should stare at a blank wall and reconsider (or, the conscious grid). Shtetl-Optimized: The Blog of Scott Aaronson.

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere?. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1668), 20140167.

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.

Tsuchiya, N. (2017). "What is it like to be a bat?"—a pathway to the answer from the integrated information theory. Philosophy Compass, 12(3), e12407.

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433-460.

Tye, M. (1995). Ten Problems of Consciousness. Cambridge, MA: MIT Press.

Tye, M. (2000). Consciousness, Color, and Content. Cambridge, MA: MIT Press.

Tye, M. (2002). Representationalism and the Transparency of Experience. Noûs, 36(1), 137-151.

Tye, M. (2018). "Qualia", The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), Edward N. Zalta (ed.)

Varela, F., & Maturana, H. (1980). Autopoiesis and Cognition: The realization of the Living.

Varela, F. J., Rosch, E., & Thompson, E. (1993). The embodied mind.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. Journal of consciousness studies, 3(4), 330-349.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks

Weiskrantz, L. (1986). Blindsight: A Case Study and Implications, Oxford: Clarendon Press.

Wetzel, L. (2018). "Types and Tokens", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.)

Yablo, S. (1993). Is conceivability a guide to possibility?. Philosophy and Phenomenological Research, 53(1), 1-42.

Yu, Q., Yang, Y., Liu, F., Song, Y. Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: A deep neural network that beats humans. International journal of computer vision, 122(3), 411-425.

Zahavi, D. (2018). Phenomenology: the basics. Routledge.