Predictive Processing and the Epistemological Hypothesis: Solving the Hard Problem of Consciousness by Simulating a Brain Facing It

Philippe Servajean[1,2,*], Richard Servajean[3,4,a]

1 EPSYLON EA, Department of Psychology, Paul-Valéry University, Montpellier, France
2 Laboratory of Psychology and NeuroCognition, University of Grenoble Alpes, France
3 Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
4 SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

*Email: philippe_gi@hotmail.fr

Abstract

When we say that a theory is able to account for our subjective experience, we simply mean that if this theory were true—e.g., if our brain were as this theory stipulates—then our subjective experience would be indeed as we experience it. A scientist would formulate this idea using the notions of *observation* and *prediction*: A theory is able to account for our subjective experience if and only if *it is able to predict first-person observations*. Several thought experiments suggest that current theories are unable to make such predictions. For example, if we had never seen colors in our lives and therefore did not know what the color blue looks like, these theories would not allow us to deduce (i.e., predict) what it is like to see blue. This well-known problem is often called the *hard problem of consciousness* (HPC). Here, we address this problem through the lens of the *epistemological hypothesis*. Under the epistemological hypothesis, the HPC no longer reflects the inability of our theories to predict first-person observations; it reflects *our* inability to deduce what these theories imply regarding first-person observations. The HPC then becomes an epistemological problem that can be formulated as follows: If we are unable to deduce what a theory implies regarding first-person observations, how can we know whether this theory is able to account for first-person observations? In this paper, we outline a method to test experimentally the epistemological hypothesis and to solve the HPC. Notably, this method makes it possible to test any *identity hypothesis*. We then highlight the remarkable compatibility between this method and the theoretical framework of *predictive processing*. We show that a theory of consciousness based on predictive processing implies the epistemological hypothesis—this theory predicts that we are unable to deduce its own implications regarding first-person observations. Finally, this work suggests that the theoretical framework of predictive processing may already have the resources to simulate a brain facing the HPC.

Keywords: hard problem of consciousness, epistemological hypothesis, type-B materialism, meta-problem of consciousness, identity hypothesis, predictive processing, precision, quality space

[a]Current address: Brain Intelligence Theory Unit, RIKEN Center for Brain Science, Saitama, Japan.

Contents

1 Introduction

There is still no consensus regarding the problem of consciousness (i.e., phenomenal consciousness, first-person observations, phenomenology, mental life, inner world, lived experience, manifest reality, conscious experience, what-it-is-like-ness [Nagel, 1974], qualia [Lewis, 1929], or subjective experience). When we say that a theory is able to account for our subjective experience, we simply mean that if this theory were true—e.g., if our brain were as this theory says—then our subjective experience would be indeed as we "see" it "from the inside" (i.e., as we experience it). In other words, we mean that this theory *necessarily implies* that our subjective experience is as it is. A scientist would formulate this idea using the notions of *observation* and *prediction*: A theory is able to account for our subjective experience if and only if *it is able to predict first-person observations*. Several thought experiments suggest that current theories are unable to make such predictions. For example, if we had never seen colors in our lives and therefore did not know what the color blue looks like, these theories would not allow us to deduce (i.e., predict) what it is like to see blue. This well-known problem is often called the *hard problem of consciousness* (HPC) (D. J. Chalmers, 1995, 1996, 2010; see also the notion of "explanatory gap" in Levine, 1983).

When confronted with the HPC, a common reaction is to conclude that current theories are unable to account for our subjective experience. However, there is another way of thinking about the HPC. Many neuroscientists and philosophers would probably agree with the following two statements:

> **1.** What happens in our brain at a given moment *necessarily implies* that our subjective experience is exactly as we "see" it at this moment (e.g., blueness).
>
> **2.** If we had never seen colors in our lives, knowing *everything there is to know* about what happens in our brain when we see the color blue would *not* allow us to deduce (i.e., predict) what it is like to see blue.

Taken together, these two statements *necessarily* mean that we are unable to deduce what a specific physical or functional state of the brain implies about first-person observations. Put another way, if we accept simultaneously these two statements, then we must consider that the HPC does not reflect the inability of our theories to predict first-person observations; it reflects *our* inability to deduce what these theories imply regarding first-person observations. This is what we have called the *epistemological hypothesis* (philosophers would speak instead of *type-B materialism* [D. J. Chalmers, 2003] or at least of some versions of type-B materialism). The basic idea behind this hypothesis is that the HPC is an epistemological problem—not an explanatory problem. If the problem of consciousness is "hard", this would not be because it is hard to develop a satisfactory theory of consciousness but rather because it is hard to *show* that this theory is satisfactory. Indeed, how could we know whether a theory is able to account for first-person observations if we are unable to deduce what this theory implies regarding first-person observations? This reading of the HPC therefore makes plausible the hypothesis that we already have a relatively satisfactory theory of consciousness without being able to show it. *There will be no consensus regarding the problem of consciousness as long as we do not have both a theory able to account for consciousness and good reasons to think that this theory is able to account for consciousness.*

The aim of the current paper is to answer two questions. First, how can we test experimentally the epistemological hypothesis? Second, under the epistemological hypothesis,

how can we solve the HPC? Our argument comprises four big steps. First, using two well-known thought experiments, we highlight the apparent inability of current theories to account for first-person observations. Importantly, these thought experiments are formulated in a particular way, and this plays a key role in our argument. Second, we show that there are in fact two ways of interpreting these thought experiments: (1) Current theories are unable to account for first-person observations; (2) *We* are unable to deduce what these theories imply regarding first-person observations; remember that the second interpretation gives rise to an epistemological problem. Third, we propose a method allowing both to test experimentally the epistemological hypothesis and to solve the HPC (as an epistemological problem). Fourth, we highlight the remarkable compatibility between this method and the theoretical framework of *predictive processing* (Friston, 2010).

2 The hard problem of consciousness

2.1 The neuroscience of consciousness

Neuroscientists test their theories using neurophysiological and behavioral measurements. These measurements are commonly referred to as *third-person observations* (i.e., "seen from the outside"). If our ultimate goal is simply to give an exhaustive account of these third-person observations, then we are not concerned with the problem in question here. Such a goal falls within the scope of the *easy* problem of consciousness[1] (D. J. Chalmers, 1995, 1996, 2010). The problem we are interested in here—the *hard* problem of consciousness—would only arise when we deal with so-called *first-person observations* (i.e., those "seen from the inside"). Most neuroscientists believe that if we have a subjective experience, and if this subjective experience is the way it is, it is no more and no less because we have a brain that possesses certain properties and/or functions in a certain way (Dehaene, 2014). This assumption has led some of these neuroscientists to set themselves the explicit goal of accounting for our subjective experience, giving rise to a new research field: the *neuroscience of consciousness* (Dehaene and Naccache, 2001). In practice, the theories emerging from this new research field are also tested using neurophysiological and behavioral measurements. However, what really matters in this case is not behavioral measurements per se, but rather what we can *infer* from these behavioral measurements.[2] Indeed, when someone tells us "I see the color blue", we are dealing with a simple verbal behavior—a third-person observation. The key point is that we can reasonably infer from this verbal behavior that this person is currently having the subjective experience of seeing blue. So, if we want to assess the ability of a theory, say Theory A, to account for first-person observations, we can simply compare the result of this inference to the predictions of Theory A:

1. According to Theory A, this person should have the subjective experience of seeing blue.

[1] Note that one may use the plural and refer to the easy problems of consciousness which would correspond to all the "usual" problems scientists deal with. The singular form, which is equivalent, would designate the problem of accounting for all the aforementioned problems.

[2] For a description of the standard experimental setup used to test theories of consciousness, see Kleiner and Hoel (2021). It has been shown that at least some theories of consciousness cannot be falsified using such a methodology (Doerig et al., 2019; Kleiner and Hoel, 2021).

2. I *infer* from the behavior of this person that he or she is currently having the subjective experience of seeing blue.

3. The prediction of Theory A is therefore correct.

It is clear that the purpose of this methodology is to assess what we can call for now the "level of fit" between first-person observations and the predictions of our theories regarding first-person observations (i.e., how much there is an agreement). It is a matter of answering the following question: *Does this theory imply that first-person observations are indeed as they are?* It is precisely when we ask ourselves this question that the HPC comes into play.[3]

Under the materialist assumption, a brain theory is supposed to be sufficient to account for consciousness. Let us consider an example of how a brain theory could enable us to anticipate (i.e., to predict) our own first-person observations. Imagine that a brilliant neuroscientist asserts to us that Theory A (i.e., a theory describing how our brain works) is able to account for our subjective experience. Then, he invites us to take part in an experiment. All we have to do is enter a room referred to as the Experimental Room. Before starting the experiment, the neuroscientist gives us two pieces of information. First, according to Theory A, our brain will be induced into state X when we are in the Experimental Room. Second, when our brain is in state X, we see the color blue. On the basis of these two pieces of information, we then formulate the following prediction: "When I will be in the Experimental Room, I will see the color blue." Finally, we carry out the experiment and find that this prediction was correct: When we were in the Experimental Room, we saw the color blue.

This example gives us an insight into how a simple brain theory could enable us to anticipate first-person observations (at least in principle). As we can see, this predictive power stems from the following two statements: (1) Theory A allows us to predict the future states of our brain; (2) A given state of our brain is always associated with the same subjective experience (e.g., state X is always associated with the subjective experience of seeing blue[4]). Note that, in practice, within the neuroscience of consciousness, the kind of predictions made, and the way these predictions are deduced, differ significantly from our illustrative example.[5]

[3]The main purpose of the neuroscience of consciousness is to account for first-person observations. However, some neuroscientists and philosophers would not agree with this statement. For example, proponents of strong illusionism (Frankish, 2016; Kammerer, 2021) would argue that (phenomenal) consciousness does not exist (in any way) and that the purpose of the neuroscience of consciousness is only to account for some specific functions of the brain and/or for some specific behavioral and neurophysiological measurements. In brief, a proponent of strong illusionism would argue that there is no such problem as the HPC.

[4]Note that state X may not refer to a single brain state but rather to a set of brain states sharing the same coordinates within a subspace of the neural state space (Fleming and Shea, 2024; Vishne et al., 2023).

[5]In practice, materialist theories of consciousness are not (or not only) brain theories. More precisely, almost all these theories also comprise a hypothesis specifying what consciousness *is* in the brain (i.e., an *identity hypothesis*) or at least a hypothesis about what neural process, function, or mechanism is associated with or is sufficient for consciousness (see the notion of *neural correlates of consciousness* [D. J. Chalmers, 2000; Koch et al., 2016]). When it comes to formulating predictions, this hypothesis plays a crucial role. The starting point is always to assume that this hypothesis is true (e.g., "if this hypothesis were true, then we should expect to observe this or that subjective experience when our brain is in this or that state" [see Kleiner and Hoel, 2021]). In contrast, in our illustrative example, we used a simple brain theory: that is, a theory that does not comprise a hypothesis about what *is* consciousness in this context.

Regardless, the fact that Theory A allowed us to anticipate the first-person observations we sampled when we were in the Experimental Room leads us to believe that Theory A is indeed able to account for these observations. However, after discussion with a philosopher friend, we decide to question the "real" nature of this predictive power with two thought experiments.

2.2 The Mary's room thought experiment

Our first thought experiment is derived from the so-called *Mary's room* thought experiment (Jackson, 1982, 1986). Here, we simply ask ourselves: If we had never seen colors in our life (e.g., we have always worn a device that converted incoming light into black-and-white images), would Theory A have allowed us to anticipate the first-person observations we sampled in the Experimental Room? Remember that we have deduced the implications of Theory A regarding these observations based on the following two pieces of information (see Section 2.1):

1. According to Theory A, our brain will be induced into state X when we are in the Experimental Room.

2. When our brain is in state X, we see the color blue.

These two pieces of information allowed us to make the following prediction: "When I will be in the Experimental Room, I will see the color blue." However, if we had never seen the color blue in our life, we would not know what the color blue looks like. Knowing that we are going to see a color labeled "blue" tells us nothing about the appearance of that color. Consequently, we would be unable to translate the sentence "I will see the color blue" into an actual concrete expectation regarding first-person observations. In short, if we had never seen colors in our life, Theory A would not have allowed us to anticipate the first-person observations we sampled in the Experimental Room. This prediction, thought to be inherent in Theory A, was in fact partly based on our prior knowledge regarding the appearance of blue.

Presumably, the conclusion of this thought experiment is that Theory A is unable to account for the appearance of blue (i.e., the subjective experience of blue). This conclusion rests on the idea that, when a theory is able to account for a certain observation, we can use this theory to know in advance the result of this observation. Put another way, if Theory A had been able to account for the appearance of blue, we could have known what it is like to see blue simply by deducing what Theory A implies about it, even if we had never seen the color blue in the past. Note that the ideas developed in this section and the next will be detailed and clarified later on.

2.3 The inverted spectrum thought experiment

Our second thought experiment is derived from the so-called *inverted spectrum* thought experiment (Locke, 1847; Shoemaker, 1982).

The fact that a theory is able to account for an observation presupposes that if this observation had been different, then *this theory would be confronted with prediction errors*. Let us consider a concrete example. Imagine that a theory implies that the appearance of blue is indeed as we "see" it. This presupposes that if the appearance of blue had been different, then this theory would have faced a *prediction error*: that is, there would be a gap between the appearance of blue and the implications of this theory regarding the

appearance of blue. From an epistemological point of view, it is precisely when a theory is confronted with such a prediction error that we are prompted to think that this theory is false, at least in some sense, and needs to be updated or abandoned. A philosopher would formulate this idea using the notion of *metaphysical impossibility*: If a theory T is able to account for an observation O, then, in every possible world where T is true, O must be observed[6]. In short, a world in which T is true and in which O is not observed is *metaphysically impossible* (see D. J. Chalmers, 2002).

This brings us to our second thought experiment: Imagine that the appearance of blue had been that of green (and vice versa). A question then arises: In this "inverted" world, would we have noticed a gap between the appearance of blue and the predictions of Theory A? To answer this question, let us take another look at the Experimental Room. As in the real world, we would have predicted: "According to Theory A, when I will be in the Experimental Room, I will see the color blue." However, unlike the real world, in the Experimental Room, we would not have seen blue but green (i.e., we would not have experienced "blueness" but "greenness"). The question is: Does this mean we would have noticed a gap between the color we perceived ("greenness") and the prediction of Theory A (blue)? The answer is no (this answer can be derived theoretically and tested; especially, see Section 6.3). Indeed, since the beginning of our life, the appearance of blue would have always been that of green. In this "inverted" world, the color we would call "blue" would therefore literally look like the color we call "green" in the real world. As a consequence, the prediction "I will see the color blue" would be translated for us into the expectation of a color having the appearance of green (i.e., into the expectation of "greenness"). In short, we would *not* have noticed a gap between the color we perceived and the prediction of Theory A. The appearances of blue and green could have been inverted, we would not have perceived any contradiction between these observations and the predictions of Theory A. A philosopher would formulate this idea using the notion of *conceivability*: A world in which Theory A is true and in which the appearances of blue and green are inverted is *conceivable*[7].

Presumably, the conclusion of this second thought experiment is again that Theory A is unable to account for the appearance of blue. As the first, this second thought experiment enables us to realize that the predictive power of Theory A was in fact partly based on our prior knowledge regarding the appearance of blue—Theory A *alone* would not have allowed us to anticipate the first-person observations we sampled in the Experimental Room.

Before moving on, we need to address something. Several empirical studies suggest that the extent to which two subjective experiences (e.g., of color) differ depends on the extent to which the brain states underlying them "differ", and especially on the distance between these brain states within a subspace of the neural state space (Broday-Dvir et al., 2023; Fleming and Shea, 2024; Malach, 2021; I. A. Rosenthal et al., 2021; Vishne et al., 2023; see Footnote 4). Hence, based on the relative differences between state X and other brain states, one could make predictions of the following form: "The subjective experience of blue should be closer to that of purple than to that of red." The issue is that, when the appearances of blue and green are inverted, this prediction is no longer

[6]It should be noted that by "an observation O", we are referring to an observation made under specific conditions C. Hence, when we say "O must be observed", we mean "O must be observed *each time C is satisfied*".

[7]Note that the notion of conceivability in general, and whether or not it entails possibility, is highly debated (Hill, 2016), especially in the context of consciousness (D. J. Chalmers, 2002; Kripke, 1980).

true: In the inverted world, the appearance of blue (i.e., greenness) would no longer be closer to that of purple than to that of red. Consequently, in this inverted world, we could have noticed a gap between the appearance of blue and the predictions of Theory A simply by comparing this appearance with that of other colors.

To avoid this objection, it can be useful to consider a more radical version of the inverted spectrum thought experiment. A version in which the appearances of *all* colors are inverted in a way that preserves, as much as possible[8], the relative differences between these appearances. For instance, while the appearance of all colors would be different from that of the real world, the (new) appearance of blue would still be closer to the (new) appearance of purple than to the (new) appearance of red. Crucially, in such an inverted world, it would be (almost) impossible to notice a gap between the appearance of blue and the predictions of Theory A.

Regardless, this enables us to clarify the following point: When we argue that "Theory A does not allow us to predict the appearance of blue" or that "Theory A cannot be falsified via the appearance of blue", we are only referring to the appearance of blue *in itself* and not to the relative difference between this appearance and that of other colors. As we shall see in Section 6.5, this nuance is critical.

2.4 Conclusion

In Sections 2.2 and 2.3, we have reviewed two thought experiments (we refer to them as the *HPC thought experiments*). These thought experiments suggest that Theory A is unable to account for our subjective experience of blue. Crucially, while we focused on the subjective experience of blue, this problem applies to *all* first-person observations. For example, using another well-known thought experiment—the philosophical zombie thought experiment—one could have put into perspective the apparent inability of Theory A to account for the very *existence* of our subjective experience (D. J. Chalmers, 1996; Kirk, 2003).

Note that Theory A does not refer to a specific existing theory of the brain. Furthermore, in practice, materialist theories of consciousness are not (or not only) brain theories (see Footnote 5). Almost all these theories also comprise a hypothesis specifying what consciousness *is* in the brain or at least a hypothesis about what aspect (e.g., mechanism, process or function) of the brain is associated with or is sufficient for consciousness (for a non-exhaustive overview of current materialist theories of consciousness, see Table 1 of Seth and Bayne, 2022). In any case, to the best of our knowledge, no existing theory of the brain and/or consciousness would have enabled us to overcome "head-on" the problem outlined in the previous sections. From now on, we will use the expression "Theory A" as a generic term referring to any materialist theory of the brain and/or consciousness, until mentioned otherwise.

So far, we have not given a precise definition to the HPC. The key point is that the HPC thought experiments give us the intuition that it is *hard*, if not impossible, to account

[8] If we specify "as much as possible", it is because (perceived) color space displays asymmetries (Byrne, 2004; Hardin, 1988; Hilbert and Kalderon, 2000; Palmer, 1999). In this context, there is no inversion scenario in which the relative differences between the appearances of colors would be *perfectly* preserved. It is worth noting that the structuralist methodology could be very useful to identify the inversion scenarios that best preserve these relative differences. In particular, the structuralist methodology consists in measuring and then modeling the relative differences between subjective experiences (e.g., of color) using a mathematical structure known as *quality space* (Clark, 1996; Cohen et al., 2015; Lee, 2021; Nosofsky, 1992; Roads and Love, 2024; D. Rosenthal, 2010; Tsuchiya and Saigo, 2021).

for our subjective experience. It is precisely from this apparent "hardness" that the HPC takes its name. The usual formulation of the HPC frames it as the problem of *explaining* our subjective experience: How does the physical brain give rise to subjective experience (D. J. Chalmers, 1995)? The issue is that, when we define the HPC in that way, we implicitly presuppose a particular interpretation of the HPC thought experiments. In the next section (3), we clarify this issue by discussing the different possible interpretations. At this point, we need a more high-level definition: *The HPC is simply the problem we face when we conduct the HPC thought experiments.*

3 The epistemological hypothesis

As mentioned in the previous section, the conclusion of the HPC thought experiments seems to be that Theory A is unable to account for first-person observations. The aim of this section is to show that there is in fact another way of interpreting these thought experiments. In practice, we will only reformulate well-known ideas in the field of philosophy of mind.

First of all, note that when we say that Theory A is unable to account for first-person observations, we are merely stating a hypothesis. To avoid trapping ourselves from the outset, we need to return to the initial observation that led us to put forward this hypothesis. This initial observation is simple: The HPC thought experiments only made us realize that Theory A does not enable us to predict first-person observations. Hypothesizing that Theory A is unable to account for first-person observations is therefore a way to explain *why* Theory A does not enable us to predict these observations. A philosopher would formulate this idea using the notions of *epistemic gap* and *explanatory gap*. In the context of Theory A, these notions can be defined as follows: (1) The notion of epistemic gap refers to the fact that Theory A does not allow us to predict first-person observations; (2) The notion of explanatory gap refers to the fact that Theory A is unable to account for first-person observations (i.e., Theory A does not imply that first-person observations are as they are). In other words, hypothesizing the existence of an explanatory gap is a way to explain why there is an epistemic gap (a discussion of a similar inference can be found in D. J. Chalmers, 2003, 2006[9]). However, once again, the HPC thought experiments only made us realize that there is an epistemic gap between Theory A and first-person observations.

Crucially, this hypothesis—there is an explanatory gap—is not the only possible way out. In fact, two hypotheses are consistent with the initial observation of the HPC thought experiments:

- The *explanatory* hypothesis: Theory A is unable to account for first-person observations (i.e., there is an *explanatory* gap).

- The *epistemological* hypothesis: *We* are unable to deduce what Theory A implies regarding first-person observations.

[9] According to Chalmers and others (D. J. Chalmers, 2003, 2006), arguments against materialism always involve the following inference: If there is an epistemic gap, it is because there is an ontological gap (i.e., because materialism is false). This inference shares close links with the one we have described in the main text. Indeed, if we consider that "Theory A" refers to all materialist theories—not only to current materialist theories—then inferring that Theory A is unable to account for consciousness is the same as inferring that materialism is false (see question 3 in Figure 1)

Both of these hypotheses indeed imply that Theory A does not allow us to predict first-person observations. The epistemological hypothesis leads to what philosophers call *type-B materialism* or *a posteriori physicalism* (D. J. Chalmers, 2003, 2006; Speaks, 2018). The purpose of type-B materialism is to make materialism compatible with the HPC thought experiments. As we have just seen, according to the explanatory hypothesis, if there is an epistemic gap between Theory A and first-person observations, it is because there is an explanatory gap between Theory A and first-person observations. This way of thinking has led some philosophers to adopt an even more radical stance: *materialism is false* (D. J. Chalmers, 1995, 1996; Jackson, 1982, 1986; Levine, 1983). In fact, this stance logically follows from the explanatory hypothesis as long as we consider that all materialist theories—not only current materialist theories—are concerned by the epistemic gap. A type-B materialist would agree with this statement: There is indeed an epistemic gap between *all* materialist theories and first-person observations: that is, even a *perfect* brain theory would be concerned by this epistemic gap. However, when it comes to explain *why* there is an epistemic gap, a type-B materialist would reject the explanatory hypothesis in favor of the epistemological hypothesis. As a consequence, for a type-B materialist, the fact that *all* materialist theories are concerned by the epistemic gap *does not mean that materialism is false* (this view on type-B materialism is derived from D. J. Chalmers, 2003, although some differences may exist). Overall, as shown in Figure 1, it is possible to classify all the existing interpretations of the HPC thought experiments using three questions: "Is there an epistemic gap?", "Why is there an epistemic gap?" and "Why is there an explanatory gap?".

What is the correct interpretation of the HPC thought experiments—the explanatory hypothesis or the epistemological hypothesis? The very nature of the HPC and the form of its solution depend on the answer to this question. Hence, we must examine it closely. First of all, we need to point out that this question only makes sense within the framework of a *realistic* conception of science (A. Chalmers, 2013). It is a matter of considering that a theory is not only an instrument that enables us to predict observations, but that it is also, in some sense, a more or less precise approximation of reality. In this context, it becomes possible to distinguish the "apparent" predictive power of a theory from its "real" predictive power (this distinction will be expanded upon later on):

- The *apparent predictive power* of a theory is simply the predictive power, in the usual sense (i.e., the ability to anticipate observations), that this theory gives *us* when we *know* it.[10]

- The *real predictive power* of a theory corresponds to the "level of fit" between observations and the implications of this theory regarding these observations. Crucially, by "implications of this theory regarding these observations", we mean "how would these observations be if this theory were true".

In principle, when we test a theory, the question we want to address is not, in itself, "would this theory allow me to predict this observation?" but rather "If the world were as this theory stipulates, would this observation be as it is?". In other words, what we want to assess is not the apparent predictive power of this theory but rather its real predictive

[10]Note that we ignore all the subtleties that might arise when we say that an agent *knows* a theory. In general, the brain of a scientist does not encode everything there is to know about the subject he is specialist of; he takes notes, consults a computer, performs simulations, etc. See the notion of *extended mind* in Clark and Chalmers (1998).

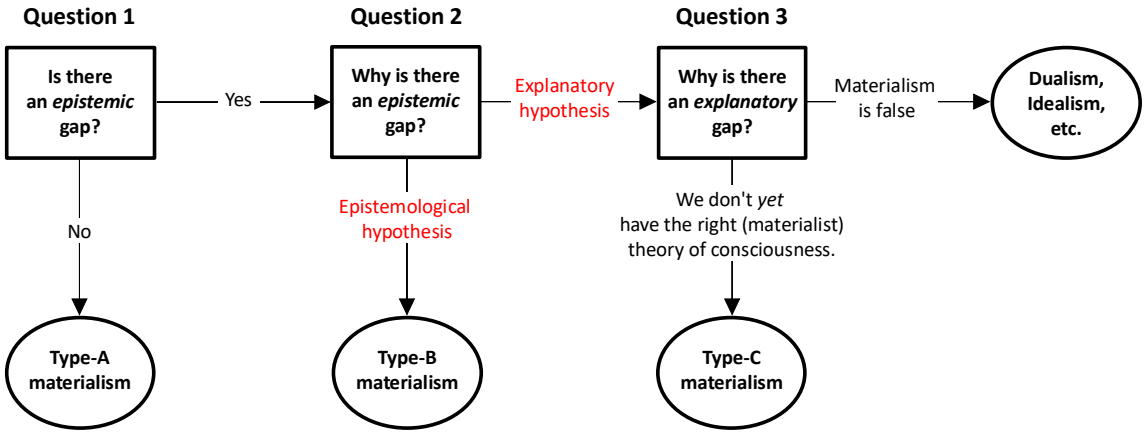

Figure 1. *Taxonomy of the possible interpretations of the HPC thought experiments.* It is possible to classify the existing interpretations of the HPC thought experiments into categories using the following three questions: "Is there an epistemic gap?", "Why is there an epistemic gap?" and "Why is there an explanatory gap?". The categories are: type-A materialism, type-B materialism, type-C materialism and non-materialist positions. This taxonomy is derived from D. J. Chalmers, 2003, although some differences may exist (see also Speaks (2018) for a brief overview). It is worth noting that most viewpoints about the nature of consciousness, the nature of the HPC or how to solve the HPC presuppose (implicitly or explicitly) a particular interpretation of the HPC thought experiments. Consequently, these viewpoints can be classified using the method presented here. As an example, *Strong illusionism* (Frankish, 2016; Kammerer, 2021), the *phenomenal concept strategy* (D. J. Chalmers, 2006; Papineau, 2002; Stoljar, 2005) and *materialist mysterianism* (McGinn, 1989) are examples of type-A, type-B and type-C materialisms, respectively.

power, unless we explicitly reject scientific realism. When we say that a theory is able to account for an observation, we are therefore referring to the real predictive power of this theory: We only mean that, if the world were as this theory stipulates, then this observation would be indeed as it is.

The issue is that only the apparent predictive power can be directly assessed. We can only assess what a theory allows us to predict. Crucially, this means that fundamental research can only be meaningful if we start from the principle that the apparent predictive power of a theory is merely a reflection of its real predictive power. The direct consequence of this principle is that, by evaluating the apparent predictive power of a theory, we indirectly evaluate its real predictive power.

It is precisely this principle that is at stake in the question "what is the correct interpretation of the HPC thought experiments—the explanatory hypothesis or the epistemological hypothesis?". Thus, we are literally wondering whether or not the apparent predictive power of Theory A reflects its real predictive power when we are dealing with first-person observations. On the one hand, the explanatory hypothesis relies on the assumption that the apparent predictive power of Theory A reflects its real predictive power. That is to say, the fact that Theory A does not allow us to predict first-person observations (apparent predictive power) reflects the fact that Theory A is unable to account for these observations (real predictive power). On the other hand, the epistemological hypothesis explicitly states that this assumption is false. Indeed, what enables the real predictive power of a theory to be translated into its apparent predictive power is precisely

our ability to deduce the implications of that theory. For instance, it goes without saying that, if we are unable to deduce what Theory A implies regarding the appearance of blue, then Theory A could have implied that the appearance of blue is as it is (real predictive power), without allowing us to predict the appearance of blue (apparent predictive power).

Here, we have just reformulated a well-known problem in the field of philosophy of mind. Indeed, with regard to the distinction between the apparent and real predictive powers, philosophers would speak instead of the distinction between *conceivability* and *metaphysical possibility*, or between *epistemic entailment* and *ontological entailment*. In the same way, when it comes to the question "does the apparent predictive power of Theory A reflect its real predictive power?", philosophers would ask "Does conceivability entail metaphysical possibility?" or "does a failure in epistemic entailment necessarily reflect a failure in ontological entailment?" (see e.g., D. J. Chalmers, 2003).

Crucially, if the epistemological hypothesis were true, we would inevitably be confronted with the following (epistemological) problem: How can we know whether a theory is able to account for first-person observations if we are unable to deduce what this theory implies regarding these observations? For example, how could we know whether Theory A is able to account for the appearance of blue if we are unable to deduce what Theory A implies regarding the appearance of blue? The epistemological tension at play is the following: If the apparent predictive power of a theory does not reflect its real predictive power, then we can no longer *infer* the "real" predictive power of this theory from its "apparent" predictive power. In this context, we would be deprived of the very source of our reasons to think that a theory is able (or unable) to account for an observation.

Finally, it is crucial to understand that there will be no consensus regarding the problem of consciousness if we do not simultaneously have a theory able to account for consciousness and *good reasons to think* that this theory is able to account for consciousness.

4 Solving the hard problem of consciousness

In this section, we propose a first approximation of a method allowing us to both test experimentally the epistemological hypothesis and to solve the HPC (as an epistemological problem). Importantly, some of the concepts developed in this section will only take on their full meaning in the next sections when we will implement this method in the theoretical framework of predictive processing.

Let us start with the following question: How could we demonstrate that the epistemological hypothesis is true? When we deduce something, we do it through our brain. That is to say, the process of deduction is a cognitive process implemented by our brain (for the sake of simplicity, here we ignore all the subtleties associated with the notion of *extended cognition* [Clark and Chalmers, 1998]). It follows that, if we are unable to *deduce* what our theories imply regarding first-person observations (i.e., if the epistemological hypothesis is true), it is probably *because of the way our brain works*. Presumably, this means that, in order to prove the epistemological hypothesis, all we have to do is demonstrate that the way our brain works necessarily implies the epistemological hypothesis.

This idea is not new. In fact, almost all type-B materialists argue that, if there is an epistemic gap, it is because of the way our brain or our cognition works (e.g., see related discussions on the phenomenal concept strategy in D. J. Chalmers, 2006; Papineau, 2002; Stoljar, 2005).

Hence, the following question: How could we show that a brain theory implies the epistemological hypothesis? At this stage of our reasoning, we are not in a position to answer this question. The reason for this is simple: The epistemological hypothesis explicitly refers to first-person observations. The issue is that it seems impossible to derive a statement that explicitly refers to first-person observations from a simple brain theory. How could a brain theory allow us to even mention first-person observations? In what follows, we argue that the only way to overcome this problem is to supplement our brain theory with a hypothesis specifying what consciousness *is* in the brain. We refer to such a hypothesis as an *identity hypothesis*[11].

4.1 An identity hypothesis

In this section and the following four ones, we illustrate our views using a specific identity hypothesis. Importantly, any other identity hypothesis could have been used instead.

Let us imagine that we are with one of our friends, say Paul, and we are looking at his brain. Now, consider the following identity hypothesis: *"The subjective experience of Paul is nothing more than the state of his brain"*. At the heart of this hypothesis lies two assertions: one explicit and one implicit. The explicit assertion is that when Paul speaks about his subjective experience and when we speak about the state of his brain, Paul and we are literally referring to *one and the same thing*. Put another way, under this identity hypothesis, the expressions "subjective experience of Paul" and "state of Paul's brain" refer to the same thing.

On the other hand, the implicit assertion involves the question: What is the difference between Paul "observing" his subjective experience and us observing the state of Paul's brain? Of course, in the context of our identity hypothesis, this difference cannot be the fact that Paul and we are observing two different things. So, what makes our situation different from that of Paul? The answer is that Paul and we are observing the same thing *from two different points of view (POV)*. Paul has a first-person POV, while we have a third-person POV.

This leads us to the following statement: Under our identity hypothesis, the expressions "subjective experience of Paul" and "state of Paul's brain" refer to the same thing *but presuppose a different POV on that thing*. The expression "subjective experience of Paul" presupposes a first-person POV, while the expression "state of Paul's brain" presupposes a third-person POV. The famous philosopher Gottlob Frege would say that these expressions have the same "referent" but a different "sense" (see Frege, 1892).

As a consequence, although these expressions refer to the same thing, they should not be used interchangeably. More precisely, it is only *when there is a change of POV* that we have to translate the expression "subjective experience of Paul" by the expression "state of Paul's brain" (or conversely). For example, this would be the case when Paul

[11] Here, we call "identity hypothesis" any hypothesis specifying what consciousness is in the brain. No matter what aspect of our brain is identified with consciousness. This could be a physical state, a functional state, or anything else. Many theories of consciousness involve an identity hypothesis. Let us consider a few examples. According to the mind/brain identity theory, Paul's subjective experience is identical to the physical state of his brain (e.g., see Smart, 2000). According to functionalism, Paul's subjective experience is identical to the functional state of his brain (e.g., see Block, 1982). According to the integrated information theory, Paul's subjective experience is identical to the causal properties of his brain (e.g., see Tononi et al., 2016). Finally, according to some versions of the higher-order theory consciousness, Paul's subjective experience is a higher-order representation of perceptual content (e.g., see Carruthers, 2017). For a non-exhaustive overview of current materialist theories of consciousness, see Table 1 of Seth and Bayne, 2022.

communicates with us, since Paul has a different POV from us. In this context, when Paul is speaking about his subjective experience, from our POV we must consider that he is currently referring to the state of his own brain.[12] In the same way, when we are speaking about the state of his brain, from his POV Paul must consider that we are currently referring to his subjective experience. Of course, this translation process only makes sense if our identity hypothesis is true; that is, if and only if the subjective experience of Paul is really nothing more than the state of his brain.

This idea shares close links with the notion of *conceptual dualism* which is at the heart of the phenomenal concept strategy (D. J. Chalmers, 2006; Papineau, 2002; Stoljar, 2005); see also Battaglia et al., 2025). According to conceptual dualism, two types of concepts operate in our brain: phenomenal concepts and physical concepts. Phenomenal concepts would be used (in particular) when we adopt a first-person POV, while physical concepts would be used when we adopt a third-person POV. Importantly, a conceptual dualist would also argue that some of our phenomenal and physical concepts literally refer to the same thing. As an example, under our identity hypothesis, the physical concept "state of Paul's brain" would refer to the same thing as the phenomenal concept "subjective experience of Paul". The key point is that, if some of our phenomenal and physical concepts indeed refer to the same thing, then, when we move from the first-person POV to the third-person POV (or conversely), we can logically *translate* our phenomenal concepts into physical concepts (or conversely), as explained in the previous paragraph.

Finally, this translation process is not just a matter of translating the expressions "subjective experience of Paul" and "state of Paul's brain" taken individually. If these expressions refer to the same thing, this remains true when they are found within a sentence. Put another way, this translation process could also be used to translate a whole sentence, such as the definitions of the apparent and real predictive power and, therefore, the definition of the epistemological hypothesis.

4.2 The epistemological hypothesis from the third-person POV

First of all, let us remember the original definitions of the apparent and real predictive powers (see Section 3):

- The apparent predictive power of Theory A is the predictive power that Theory A gives to Paul regarding first-person observations.
- The real predictive power of Theory A is the level of fit between first-person observations and the implications of Theory A regarding first-person observations.

As we can see, these two definitions explicitly refer to first-person observations—to Paul's subjective experience. Therefore, it is by using this terminology that *Paul* would formulate these notions. Remember that Paul has a first-person POV. A question then arises: How should we formulate these notions when we consider the situation from a third-person POV? If the subjective experience of Paul is really nothing more than the state of his brain, then the answer is straightforward:

- The apparent predictive power of Theory A is the predictive power that Theory A gives to Paul regarding the *state of his own brain*.

[12] What can be confusing here is that, whatever the POV that is physically imposed on us, we can "mentally" change our POV. For example, even when we (initially) have a third-person POV, nothing prevents us from using the expression "subjective experience of Paul". This would mean that we put ourselves in Paul's shoes and therefore changed our POV.

- The real predictive power of Theory A is the level of fit between the *state of Paul's brain* and the implications of Theory A regarding the *state of Paul's brain*.

Crucially, under our identity hypothesis, these new definitions and the original ones are strictly equivalent (i.e, refer to the same thing). For example, this means that when we say "the predictive power that Theory A gives to Paul regarding the state of his own brain" and when Paul says "the predictive power that Theory A gives me regarding first-person observations", Paul and we are literally referring to one and the same thing. Note that there is nothing speculative here. Of course, it goes without saying that our identity hypothesis could be false. However, the only thing that matters here is the logical implication: "*If* this identity hypothesis is true, *then* [...]".

Regardless, the new definition of the apparent predictive power is very surprising. Indeed, it is not about Paul trying to predict the state of his own brain in the usual sense, that is from a third-person POV. Here, it is about Paul trying to predict the state of his own brain *directly "from the inside"* (i.e. from the first-person POV). To develop further this point, we must first place ourselves within a well-defined neuroscientific theoretical framework (this new definition of the apparent predictive power will take on its full meaning when introducing the theoretical framework of predictive processing and especially in Section 6).

Finally, remember that the epistemological hypothesis is only the hypothesis stating that the apparent predictive power does not reflect the real predictive power. Hence, the simple fact of reformulating the notions of apparent and real predictive power, as we have just done, is sufficient to reformulate the epistemological hypothesis itself. In short, we now possess a "third-person" version of the epistemological hypothesis. In what follows, we refer to the two versions of the epistemological hypothesis using two different expressions. The "first-person" version is called the *subjective epistemological hypothesis* and the "third-person" version is called the *objective epistemological hypothesis*.

4.3 How can we show that a theory implies the subjective epistemological hypothesis?

Initially, we were unable to show that a brain theory implies the *subjective* epistemological hypothesis. The issue was that this version of the epistemological hypothesis explicitly refers to first-person observations. Fortunately, we now possess a version of the epistemological hypothesis that does not refer to first-person observations; this is the *objective* epistemological hypothesis (see previous section). Since it does not refer to first-person observations, it is "easy" to determine whether a brain theory implies the objective epistemological hypothesis. That is to say, this task falls within the scope of the "easy" problem of consciousness.

Furthermore, if our identity hypothesis is true, then the subjective and objective epistemological hypotheses are literally two different ways of formulating *one and the same hypothesis* (it is the same hypothesis formulated from two different POVs). In other words, under our identity hypothesis, saying that the objective epistemological hypothesis is true is the same as saying that the subjective epistemological hypothesis is true.

In summary, we have just highlighted the following two points:

1 It is "easy" to determine whether a brain theory implies the *objective* epistemological hypothesis.

2 Under our identity hypothesis, saying that the objective epistemological hypothesis is true is the same as saying that the subjective epistemological hypothesis is true.

Taken together, these two statements mean that it is "easy" to determine whether a theory implies the *subjective* epistemological hypothesis *if and only if this theory contains our identity hypothesis*. As an example, let us assume that Theory A+I is a theory made up of both Theory A (i.e., a brain theory) and our identity hypothesis. In this context, the simple fact of showing that Theory A implies the objective epistemological hypothesis (which is an "easy" task) is sufficient to demonstrate that Theory A+I implies the subjective epistemological hypothesis. The basic idea here is the following: If the fact that Theory A is true implies the fact that the objective epistemological hypothesis is true, then the fact that *both* Theory A and our identity hypothesis are true necessarily implies the fact that the subjective epistemological hypothesis is true.

In short, this reasoning would allow Paul to say: "*If* my brain works as Theory A stipulates *and if* my subjective experience is really nothing more than what this identity hypothesis stipulates, *then* the subjective epistemological hypothesis is necessarily true (i.e. I am unable to deduce what a theory implies regarding first-person observations)". Figure 2 sums up the reasoning.

In Section 6, we will consider a concrete example of how this reasoning can be implemented in practice. This implementation will enable us to show that a theory of consciousness based on the theoretical framework of predictive processing implies the subjective epistemological hypothesis. Note that what we call "theory of consciousness" is a theory made up of both a brain theory and an identity hypothesis.

Crucially, the fact of showing that a theory implies the subjective epistemological hypothesis is already an important goal. Remember that, if one tends to believe that a theory is unable to account for first-person observations, it is because this theory does not allow us to predict these observations (refer to Section 3). The simple fact of showing that this theory implies the subjective epistemological hypothesis renders this reasoning obsolete, and the reason is simple: As soon as this theory itself tells us that its apparent predictive power does not reflect its real predictive power, the inference of the real predictive power of this theory from its apparent predictive power becomes *flawed*. In short, the fact that this theory does not allow us to predict first-person observations is no longer a reason to conclude that this theory is unable to account for these observations.

4.4 Solving the meta-problem of consciousness and testing our identity hypothesis

In this section, we show that these ideas have concrete and experimentally testable implications.

Let us imagine that Paul says: "Theory A is unable to account for first-person observations". We then ask Paul: "Why do you say that Theory A is unable to account for first-person observations?". In this context, Paul would be likely to answer our question by describing the process we outlined in Section 3, namely:

- Step 1: I conducted the inverted spectrum thought experiment.
- Step 2: I realized that Theory A does not give me any predictive power regarding first-person observations.

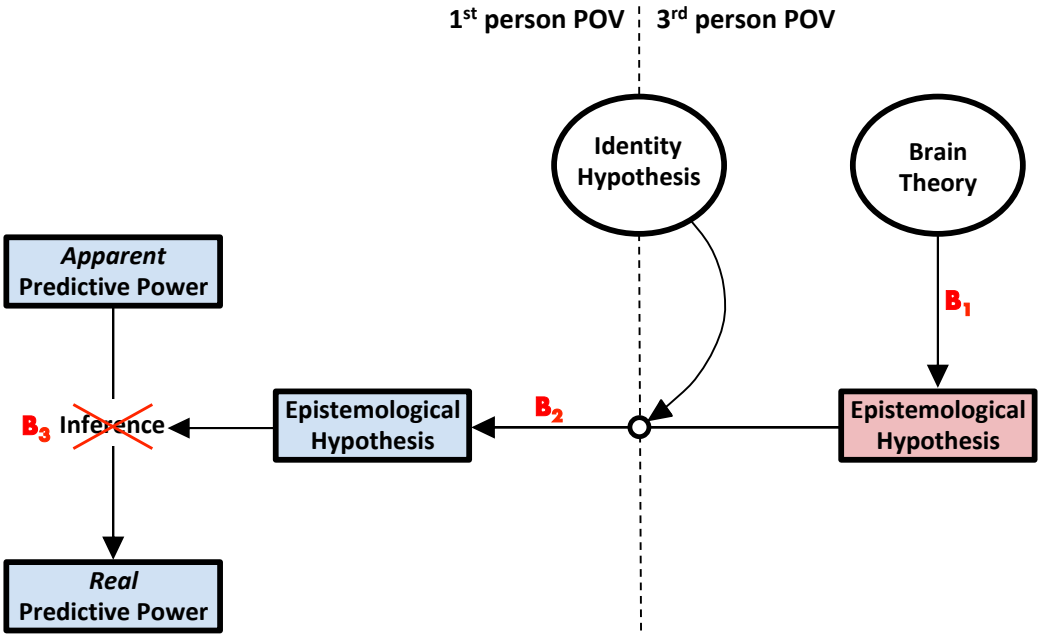

Figure 2. *Showing that a theory of consciousness implies the subjective epistemological hypothesis and updating our interpretation of the HPC thought experiments.* Remember that what we call "theory of consciousness" is a theory made up of both a brain theory and an identity hypothesis. The vertical dotted line marks the separation between the first-person POV (light blue rectangles) and the third-person POV (light red rectangle). The depicted steps, labelled B_1, B_2, B_3 (see Figure 4) consist in the following: B_1 is showing that our brain theory implies the *objective* epistemological hypothesis, B_2 consists in formulating the epistemological hypothesis from the first-person POV using our identity hypothesis and B_3 is updating our interpretation of the HPC thought experiments through the lens of the *subjective* epistemological hypothesis.

- Step 3: I put forward the hypothesis that Theory A is unable to account for first-person observations.
- Step 4: I said "Theory A is unable to account for first-person observations".

The key point here is that Paul explicitly refers to first-person observations. If Paul uses this terminology, it is because he has a first-person POV, which is not the case for us. A question then arises: How should we describe this process when we consider the situation from a third-person POV? Under our identity hypothesis, the answer is simple:

- Step 1: Paul conducted the inverted spectrum thought experiment.
- Step 2: He realized that Theory A does not give him any predictive power regarding the *state of his own brain* (remember that it is about Paul trying to predict the state of his own brain directly "from the inside").
- Step 3: He put forward the hypothesis that Theory A is unable to account for the *state of his own brain*.
- Step 4: He said "Theory A is unable to account for first-person observations".

Here, we have just reformulated the notions of apparent predictive power (see Step 2) and real predictive power (see Step 3) exactly in the same way as we did in Section 4.2. Crucially, this four-step process must be considered as a very "high-level" approximation of the *physical process* that took place within Paul's brain and that led him to the verbal behavior of saying: "Theory A is unable to account for first-person observations". A philosopher would say that describing this physical process amounts to solving what is known as the *meta-problem of consciousness* (D. J. Chalmers, 2018). Indeed, the meta-problem of consciousness can be defined as the problem of explaining why we say (and think) that materialist theories (e.g., Theory A) are unable to account for consciousness.[13]

For the moment, we only possess a very "high-level" approximation of this physical process. In what follows, we explain how we could obtain an accurate description of what happens in Paul's brain during each step of this physical process.

Let us start with the first step: "Paul conducts the inverted spectrum thought experiment". In principle, deducing what happens in Paul's brain when he performs a cognitive task, as conducting a thought experiment, falls within the scope of the easy problem of consciousness. To do so, we only need to answer the following question: *Given how it works, how could the brain of Paul implement this thought experiment?*

The issue is that the thought experiment of Paul involves *first-person observations*. The question of how Paul's brain processes information about the external world is an "easy" question. However, it seems much more difficult to answer the question of how Paul's brain processes information—and conducts thought experiments—about *first-person observations*.

If the ideas we have just developed are on the right track, then this difficulty has something to do with our POV. When we investigate how Paul's brain could implement this or that cognitive task, by definition, we are adopting a third-person POV. On the other hand, the inverted spectrum thought experiment explicitly refers to first-person observations: "Imagine a world in which the appearance of blue is that of green (and vice versa)". This means that an observer would use this terminology to describe the "inverted spectrum" scenario if and only if he were observing the situation from the first-person POV, which is not the case for us. Therefore, we need to answer the following question: How should we describe the "inverted spectrum" scenario, when we consider the situation from the third-person POV?

In its current form, our identity hypothesis is not sufficient to answer this question. The reason for this is that our identity hypothesis does not specify what the appearances of blue and green are. Hence, to pursue our reasoning, let us add a few more details to this identity hypothesis: "Blueness" and "greenness" are nothing more than state X and state Y of Paul's brain, respectively. Crucially, if this refined version of our identity hypothesis is true, then from our POV (i.e., the third-person POV), the "inverted spectrum" scenario must be described as follows: "Imagine a world in which state X systematically occurs in Paul's brain instead of state Y (and vice versa)". It is crucial to understand precisely what we mean here. Under our identity hypothesis, this new version of the inverted spectrum thought experiment and the original one are literally two different ways of describing one and the same scenario (it is the same scenario described from two different POVs). This means that when Paul conducts the inverted spectrum thought experiment, *from our POV*, we must consider that he is literally conducting the following thought experiment: "Imagine a world in which state X systematically occurs in *my* brain instead of state

[13] Alternatively, the meta-problem of consciousness can also be defined as the problem of explaining why we say (and think) that consciousness poses a "hard" problem (D. J. Chalmers, 2018).

Y (and vice versa)". Again, there is nothing speculative here, since the only thing that matters is the logical implication: "*if* our identity hypothesis is true, *then* [...]".

Let us clarify a critical detail. The new version of the inverted spectrum thought experiment does not explicitly refer to first-person observations. Hence, one could understand that Paul conducts this thought experiment from the third-person POV (i.e., as if he were observing his own brain from the outside). Crucially, this is not what we mean. Here, it is about Paul conducting this thought experiment from the first-person POV (this idea takes on its full meaning in Section 7). If we describe the thought experiment of Paul without referring to first-person observations, it is no more and no less than because *we* have a third-person POV (no matter the POV of Paul).

We were unable to answer the question: Given how it works, how could the brain of Paul implement the inverted spectrum thought experiment? The issue was that this thought experiment explicitly refers to first-person observations. Fortunately, we now possess a version of this thought experiment that does not refer to first-person observations. We can therefore directly use this new version to answer our question: *Given how it works, how could Paul's brain implement this thought experiment?* At this point, it is "easy" to tackle this question since the thought experiment at stake here no longer refers to first-person observations. In short, we are now in a position to deduce what happens in Paul's brain when he conducts the inverted spectrum thought experiment. In Section 7, we will consider a concrete example of how this reasoning can be implemented in practice by leveraging the theoretical framework of predictive processing. This will enable us to outline predictions about what happens in our brain when we conduct the inverted spectrum thought experiment and, more generally, when we face the HPC.

In summary, this reasoning makes it possible to formulate predictions about the *neural correlates of the HPC*: that is, predictions about what happens in Paul's brain when he is facing the HPC. This is therefore an opportunity to test experimentally our identity hypothesis. The reason for this is clear: These experimental predictions stem from a reasoning that is explicitly based on our identity hypothesis (see Figure 3).

Let us insist more on this point because it is a crucial step. One might have thought that an identity hypothesis could not be tested through third-person observations. Since third-person observations do not fall within the scope of the HPC, we are supposed to be able to give an exhaustive account of them without *ever* referring to the notion of subjective experience (and therefore without ever referring to our identity hypothesis). The basic idea here is that the outcome of any neurophysiological measurement is entirely determined by what is happening in Paul's brain. As a consequence, these neurophysiological measurements could *only* be used to test our hypotheses about what is happening in Paul's brain. Of course, this statement is supposed to be true even when we are dealing with neurophysiological measurements *regarding the neural correlates of the HPC*. How is it then that our identity hypothesis can be tested experimentally through neurophysiological measurements regarding the neural correlates of the HPC? The answer is simple: The core idea of our reasoning is to use our identity hypothesis *as a means* to deduce what is happening in Paul's brain when he is facing the HPC. Hence, our hypotheses about what is going on in Paul's brain when he is facing the HPC are supposed to be true *if and only if our identity hypothesis is also true*. When we test these hypotheses experimentally, we are therefore indirectly testing our identity hypothesis.

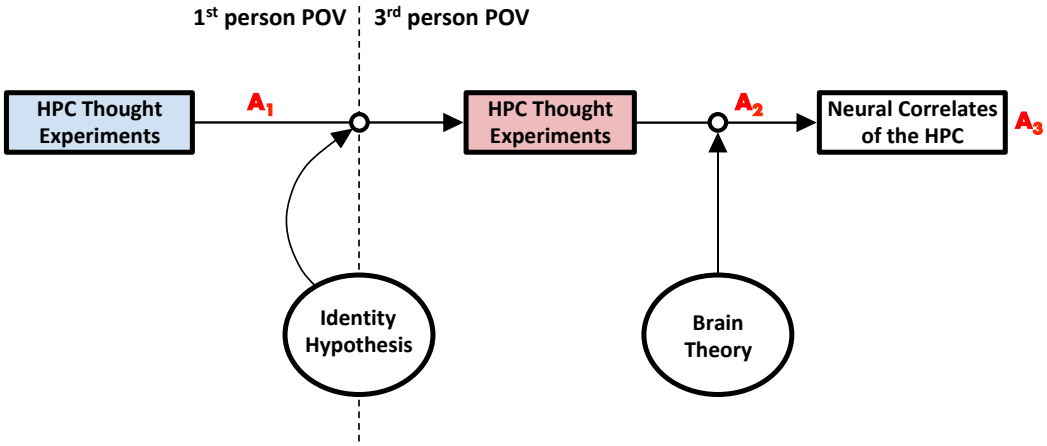

Figure 3. *Testing our identity hypothesis.* The vertical dotted line marks the separation between the first-person POV (light blue rectangle) and the third-person POV (light red rectangle). The depicted steps, labeled A_1, A_2, A_3 (see Figure 4), consist in the following: A_1 is dedicated to formulating the HPC thought experiments from the third-person POV using our identity hypothesis, A_2 consists in using our brain theory to describe how a brain could actually implement these thought experiments (i.e., to predict the neural correlates of the HPC) and A_3 is experimentally testing these predictions and concluding regarding the validity of our identity hypothesis.

4.5 Going back to the first-person POV and solving the hard problem of consciousness

A final question arises: How could this experimental test enable Paul to obtain *good reasons* to think that Theory A is able to account for first-person observations? As we have seen, our identity hypothesis gives rise to a number of equivalences between the first-person perspective and the third-person perspective (see Table 1). One of these equivalences concerns the notion of real predictive power. So, by obtaining good reasons to think that our identity hypothesis is true, Paul obtains de facto good reasons to think that our two definitions of the real predictive power (i.e., the subjective and objective definitions) are indeed strictly equivalent (i.e., refer to the same thing). In this context, *assessing the level of fit between the state of Paul's brain and the implications of Theory A regarding the state of Paul's brain, which is an "easy" problem, is the same as assessing the level of fit between first-person observations and the implications of Theory A regarding first-person observations.* Using this method, Paul could therefore answer the following question: "Does Theory A imply that first-person observations are indeed as they are?"

In short, by completing all the steps outlined in this section and the previous ones, we could both test experimentally the epistemological hypothesis[14] and solve the HPC (as an epistemological problem). Figure 4 contains a summary of all these steps. Importantly, it should be noted that some of the concepts we have developed so far have been approximately defined. One of the main purposes of the next three sections is to clarify these concepts and to propose an actual implementation of our method. To do so, we will introduce these concepts within a well-defined neuroscientific theoretical framework: *predictive processing* (Clark, 2013; Friston, 2010).

[14] Figure 2 shows that, after establishing the objective epistemological hypothesis , which is in principle an "easy" problem, experimentally validating our identity hypothesis is equivalent to experimentally

Table 1 | Equivalences induced by our identity hypothesis (this hypothesis is only used as an example)		
	First-person POV	**Third-person POV**
Identity hypothesis	My subjective experience	State of Paul's brain
***Real* predictive power**	Level of fit between first-person data and the implications of Theory A regarding first-person data	Level of fit between the state of Paul's brain and the implications of Theory A regarding the state of Paul's brain
***Apparent* predictive power**	Predictive power that Theory A gives me regarding first-person data	Predictive power that Theory A gives to Paul regarding the state of his own brain
Epistemological hypothesis	The apparent predictive power of Theory A does not reflect its real predictive power	The apparent predictive power of Theory A does not reflect its real predictive power
Question (Meta-problem)	Why do I say « Theory A is unable to account for first-person observations »?	Why does Paul say « Theory A is unable to account for first-person observations »?
Answer	I conducted the Inverted Spectrum thought experiment. ↓ Ascertainment I realized that Theory A does not allow ME to predict first-person observations. ↓ Inference I hypothesized that Theory A is unable to account for first-person observations. ↓ Behavior I said « Theory A is unable to account for first-person observations ».	Paul conducted the Inverted Spectrum thought experiment. ↓ Ascertainment He realized that Theory A does not allow HIM to predict the state of his own brain. ↓ Inference He hypothesized that Theory A is unable to account for the state of his own brain. ↓ Behavior Paul said « Theory A is unable to account for first-person observations ».

validating the (subjective) epistemological hypothesis.

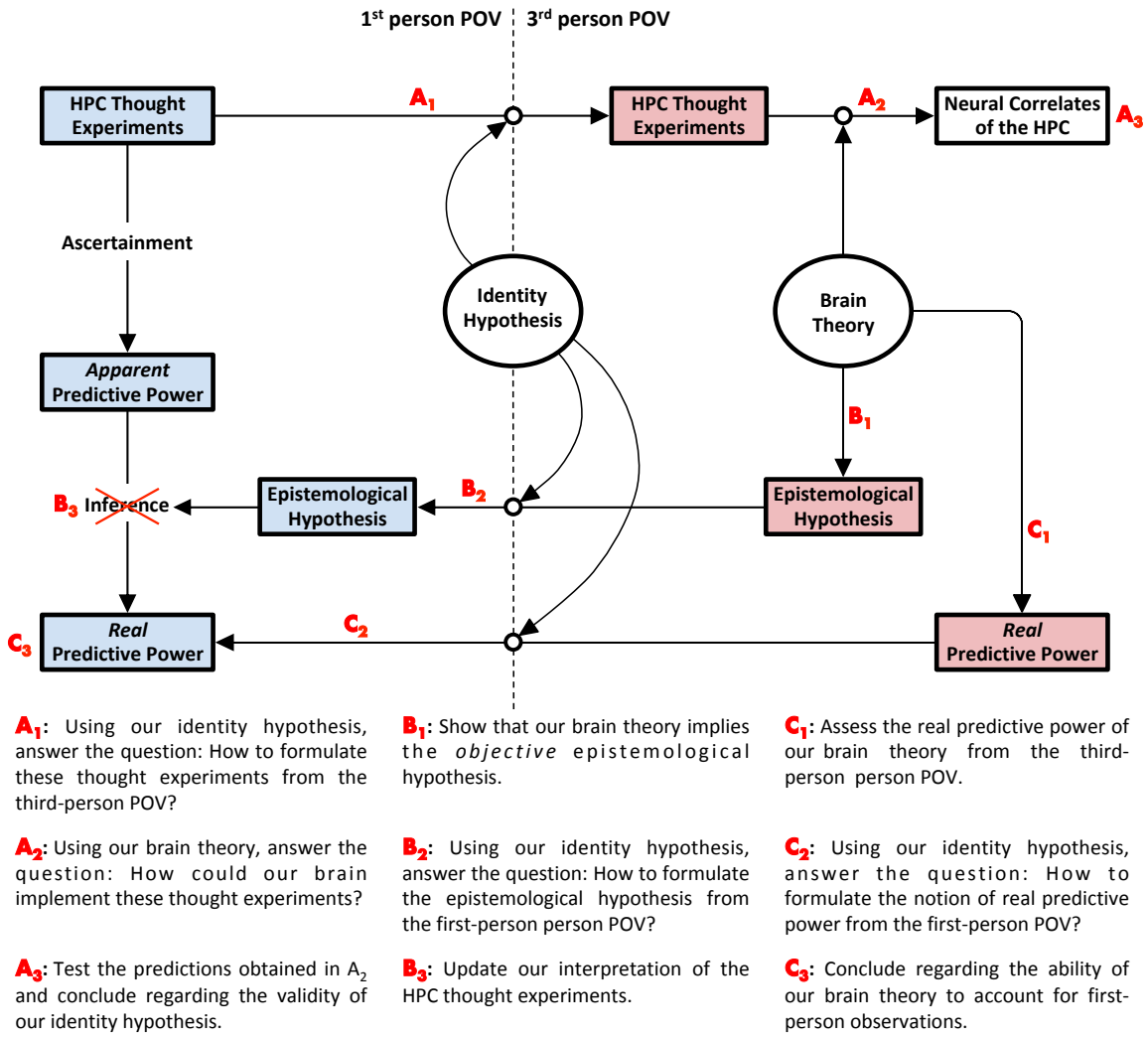

Figure 4. *Summary of all the steps that must be completed to demonstrate that a theory is able to account for first-person observations.* These steps correspond here to A_1, A_2, A_3, B_1, B_2, B_3, C_1, C_2, and C_3 (note that this order is not reflected in the way we conducted our reasoning in the main text). The vertical dotted line marks the separation between the first-person POV (light blue rectangles) and the third-person POV (light red rectangles). More precisely, a red background means that we are dealing with the objective version of a concept and a blue background means that we are dealing with its subjective version. It is possible to consider that a *theory of consciousness* (i.e. a brain theory associated with an identity hypothesis) is able to account for first-person observations if it successfully goes through all the steps involved. If a theory fails at steps B_1, B_2 or B_3, then the HPC thought experiments will continue to give us reason to think that this theory is *unable* to account for first-person observations (see especially Section 4.3). The purpose of the steps C_1, C_2, and C_3 is to assess directly the ability of our theory to account for first-person observations (see especially Section 4.5). However, these steps (i.e., C_1, C_2, and C_3) only make sense if our identity hypothesis is true. The purpose of the steps A_1, A_2 and A_3 is precisely to test this identity hypothesis (see especially Section 4.4).

5 A theory of consciousness based on predictive processing

The aim of this section is twofold. First, we briefly present the theoretical framework of predictive processing (PP). Second, we formulate an identity hypothesis under the PP framework.

5.1 The predictive processing framework

PP is a unifying brain theory: that is, a theory whose ambition is to "explain perception and action and everything mental in between"[15] (Hohwy, 2013, p. 1). Many studies have already provided evidence and arguments suggesting that PP has the resources to account for well-known cognitive functions (see Hohwy, 2020), such as, for example: perception (Parr and Friston, 2018; Rao and Ballard, 1999; van Heusden et al., 2019; Walsh et al., 2020), action (Adams et al., 2013; Friston et al., 2009; Shadmehr et al., 2010), attention (Ainley et al., 2016; Feldman and Friston, 2010; Kanai et al., 2015; Parr and Friston, 2019a), motivation (Miller Tate, 2021), emotions (Barrett, 2017; Joffily and Coricelli, 2013; Ridderinkhof, 2017; Wilkinson et al., 2019), counterfactual reasoning (Corcoran et al., 2020) and theory of mind (Friston and Frith, 2015). We will not present this framework exhaustively. Our focus is only on the aspects of PP relevant for our purpose.

The core idea of PP is that the brain infers the causes of sensory inputs and makes predictions[16] about sensory inputs given their causes (Friston, 2010; Hutchinson and Barrett, 2019; Wiese and Metzinger, 2017). The causes of sensory inputs correspond to external world states, often referred to as hidden (or latent) states (Wiese and Metzinger, 2017). To do so, the brain relies on a generative model of sensory inputs, that is a probabilistic model describing the generative process of sensory inputs (i.e., the process of hidden states causing sensory inputs). This model captures the statistical dependencies between hidden states and the sensory consequences of hidden states (Friston, 2010). Importantly, this probabilistic model is *hierarchical*, as reflected in the cortical hierarchy of the brain (Clark, 2013; Felleman and Van Essen, 1991; Friston, 2005; Maunsell and van Essen, 1983; Parr et al., 2022; Rao and Ballard, 1999).

More precisely, the machinery is as follows: Based on previously learned causal regularities, each level of the cortical hierarchy makes predictions about the activity of the level right below which provides its inputs, with the lowest level directly dealing with sensory inputs. Technically, a prediction corresponds to the mode of a belief (i.e., a probability distribution).[17] The discrepancy between predictions and what is predicted then gives rise to ascending prediction error signals used to update predictions in a Bayes optimal fashion. In other words, predictions are updated in order to minimize prediction errors. This updating process allows the brain to determine the most probable hypotheses about the causes of sensory inputs: that is, the hypotheses or predictions with the highest posterior probability. If you are not familiar with Bayes' theorem, note that the posterior

[15] Although some authors may use the names predictive processing and predictive coding in an interchangeable fashion (e.g., Cencini et al., 2021), we refer to predictive processing as a unifying brain's theory (Clark, 2013; Friston, 2010; Wiese and Metzinger, 2017) and, under this general framework, predictive coding corresponds to a computational model of perceptual inference (Rao and Ballard, 1999).

[16] The word prediction can be used to denote expectations regarding hidden states or the sensory inputs that are caused by hidden states.

[17] In the context of Gaussian beliefs, the mode and the mean are equal.

probability of a hypothesis "[...] depends on the likelihood (i.e., how well the hypothesis predicts the input); and on the prior probability of the hypothesis (i.e., how probable the hypothesis was before the input)" (Hohwy et al., 2008, p. 688).

Crucially, these hypotheses about the causes of sensory inputs are encoded across all levels of the cortical hierarchy. Indeed, when a given level makes predictions about the level below, it forms hypotheses about the external world. In fact, the different levels capture different degrees of abstraction and operate at different timescales (see the notion of deep temporal models [Friston, Rosch, et al., 2017; Parr et al., 2022]). For example, in the context of reading, while a low level would infer letters, a higher level would infer words, and so on (Friston, FitzGerald, et al., 2017; Parr et al., 2022).

The updating of predictions depends on *precision*. Formally, precision is the inverse variance of beliefs.[18] "From an empirical point of view, higher precision implies more vigorous belief updating [...]" (Parr et al., 2022, p. 157). The reason for this is that:

> "[...] the information contained in prediction errors can be more or less reliable. This reliability depends on the *cause* of prediction errors: Some prediction errors are indeed the result of inadequate predictions, while others only reflect the unpredictability of sensory signals (i.e., the large variance or low precision of sensory signals). In order to update predictions in a relevant way, it is therefore necessary to weight prediction errors according to their estimated precision; that is, according to the estimated reliability of the information they contain." (Servajean and Wiese, 2024, p. 3)

Thus, the more precise a prediction error is supposed to be given the context, the more it is weighted and the more it is likely to influence how our beliefs are updated. As such, precision estimates "have a profound effect on the subsequent inferences because they change the way that ascending prediction errors shape our beliefs" (Servajean and Wiese, 2024, p. 3). Notably, precision is often equated with attention (Ainley et al., 2016; Feldman and Friston, 2010; Kanai et al., 2015; Parr et al., 2022). Let us consider a concrete example. We find ourselves in a very familiar environment, but some fog obscures our vision. In such a situation, the estimated precision of higher and lower perceptual levels would be different. On the one hand, since we are in a very familiar environment, the precision ascribed to higher perceptual levels would be high. On the other hand, since it is very foggy, the precision ascribed to lower perceptual levels (close to our visual inputs) would be low—we would not pay attention to our visual inputs. As a consequence, our *posterior* beliefs about the external world would be based mainly on our prior knowledge rather than visual evidence (i.e., posterior expectations would be closer to prior expectations).

As we have already mentioned, predictions rest on previously learned causal regularities (FitzGerald et al., 2015). Such causal regularities are encoded by the brain via synaptic weights (Friston, 2010). Importantly, these synaptic weights are parameters of the generative model. Learning is therefore the optimization of the model parameters in light of new evidence. This is what is called *Bayesian parameter learning* (Rutar et al., 2023). Prediction errors play a key role in this process. Indeed, prediction errors are not only used to update predictions but also to update the synaptic weights that encode

[18] In active inference on discrete state-spaces (i.e., when dealing with categorical distributions represented by vectors) (Da Costa et al., 2020), precision over policies (i.e., confidence in which policies are selected) is defined as an inverse temperature (or thermodynamic β) for the softmax normalization. The idea is similar when considering precision for different parts of the generative model, as, for example, the likelihood (see Subsection B.2.4 in Parr et al., 2022).

causal regularities of the external world. Having said this, it should be noted that learning cannot be reduced to optimizing the preexisting parameters of the generative model (Rutar et al., 2022). A comprehensive account of learning must also explain how the very structure of the generative model is learned, e.g. how parameters are added. This other kind of learning is known as *structure learning* (Smith et al., 2020).

Finally, an important question arises: How can synaptic weights *in the brain* reflect causal regularities existing *in the external world*? The answer is that causal regularities existing in the external world give rise to statistical regularities in our sensory inputs and the purpose of our cortical hierarchy is to track and learn these statistical regularities. Lower levels learn statistical regularities occurring at short spatial and temporal scales, while higher levels learn statistical regularities occurring at larger spatial and temporal scales. This function is fulfilled both by structure learning and Bayesian parameter learning. In the context of PP, learning is indirect for two reasons. First, causal regularities of the external world are learned indirectly through the statistical regularities in our sensory inputs. Second, each level of the cortical hierarchy bases its learning on the level below and not directly on sensory inputs, except for the lowest level.

We are now ready to formulate an identity hypothesis in the context of PP.

5.2 The winning hypothesis theory

Keep in mind that our aim in this section is not to propose the most accurate identity hypothesis possible, but rather the simplest one that is compatible with the ideas we have previously developed.

One of the most popular views regarding consciousness in PP is that the content of consciousness is determined by the winning hypothesis (Hohwy, 2013; Hohwy et al., 2008; Ramstead, Albarracin, et al., 2023; Whyte et al., 2024), that is, the hypothesis or prediction with the highest (posterior) probability: "[...] perceptual content *is* the predictions of the currently best hypothesis about the world" (Hohwy, 2013, p. 48). Put another way, "[...] perceptual content is equated with the content of the probabilistic representations that make up the model of the world constructed by the brain" (Schlicht and Dolega, 2021, p. 18). Our identity hypothesis shares close links with this idea. More precisely, from now on we will consider that our subjective experience is nothing more than the set of all winning hypotheses in our brain. Of course, without a precise description of how neurophysiological mechanisms implement PP, this identity hypothesis does not really tell us (or at least not precisely) what consciousness is in the brain. Regardless, as we shall see, this identity hypothesis is perhaps incomplete or even false[19], but it is sufficient to highlight the remarkable compatibility between PP and the ideas we have previously developed. Henceforth, we will call the winning hypothesis theory, a theory that incorporates *both* PP and this identity hypothesis. In short, under the winning hypothesis theory, our brain is conformed to PP and our subjective experience at a given

[19]Perhaps a more plausible version of this identity hypothesis is that our subjective experience is only a *subset* of all the winning hypotheses. However, this raises the following question: What makes a winning hypothesis part of this "conscious" subset? Current theories of consciousness could be very useful to answer this question (in particular, the higher-order theories of consciousness [Carruthers, 2017], the global workspace theory [Baars, 1993, 2005; Dehaene, 2014] and the intermediate-level theory of consciousness [Jackendoff, 1987; Marchi and Hohwy, 2022; Prinz, 2017]). As an example, if we follow some versions of the higher-order theory of consciousness, we obtain the following identity hypothesis: Our subjective experience is nothing more than the set of all winning hypotheses *that are the subject of a higher-order representation (i.e. a meta-representation).*

moment is nothing more than the set of all winning hypotheses in our brain at that moment.

What does the winning hypothesis theory tell us about our subjective experience of colors? According to the winning hypothesis theory, the blueness of the sky, the redness of poppies, and the greenness of forests are nothing more than hypotheses about the causes of sensory inputs. A question then arises: Where exactly do these hypotheses arise in our cortical hierarchy? Within PP, there are important differences between higher and lower levels of this cortical hierarchy. First, as we have already mentioned, higher levels encode more abstract information than lower levels. Second, higher levels encode amodal or multimodal information while lower levels encode information regarding specific sensory modalities (Mesulam, 1998). Presumably, the hypotheses corresponding to our subjective experience of colors would belong to the visual part of our cortical hierarchy and therefore to the lower levels (Tong, 2003). For the sake of simplicity, we will assume that our subjective experiences of blue and green are two hypotheses of a categorical belief (as opposed, e.g., to a Gaussian belief) that belongs to the lowest level of our cortical hierarchy (this is a simplification for several reasons[20]). We will refer to these hypotheses using the expressions *representation of blue* and *representation of green*. In other words, when we say, e.g., "The representation of blue occurs in our brain", it should be understood that the hypothesis corresponding to our subjective experience of blue is part of the *winning* hypotheses.

6 The winning hypothesis theory and the epistemological hypothesis

The aim of this section is to demonstrate that the winning hypothesis theory implies the *subjective* epistemological hypothesis. Remember that, to achieve this goal, we only need to show that PP—alone—implies the *objective* epistemological hypothesis. This is because the objective and subjective versions of the epistemological hypothesis are strictly equivalent under our identity hypothesis and that the winning hypothesis theory incorporates *both* PP and our identity hypothesis (see Section 4.3 and Figure 2).

First of all, let us introduce the *objective* epistemological hypothesis within the theoretical framework of PP. Generally speaking, the epistemological hypothesis is simply the hypothesis stating that the apparent predictive power of a theory does not reflect its real predictive power. As discussed in Section 4, the problem is that the notions of apparent and real predictive powers are initially defined from the first-person POV:

1. The real predictive power of a theory is the level of fit between first-person observations and the implications of this theory regarding first-person observations.

2. The apparent predictive power of a theory is the predictive power that this theory provides to Paul regarding first-person observations.

[20]This is a simplification because the brain is thought to be a hierarchical model mixing discrete (i.e., categorical) and continuous variables (we speak of hybrid or mixed models) (Friston, Parr, and de Vries, 2017; Parr et al., 2022; Whyte et al., 2024). In this context, lower levels tend to be continuous (e.g., encoding luminance) while higher levels tend to be categorical (e.g., "Which country does this flag represent?") (Parr et al., 2022). Also, color space is actually a (continuous) *multidimensional* space. Notably, this suggests that a particular subjective experience of color would be in fact a combination of *several* winning hypotheses.

Therefore, we need to answer the following question: How should we define these notions when we consider the situation from the third-person POV? To do so, we are supposed to rely on our identity hypothesis. However, for the sake of simplicity, we will directly use the definitions we obtained in Section 4.2, namely:

1. The real predictive power of a theory is the level of fit between the state of Paul's brain and the implications of this theory regarding the state of Paul's brain.

2. The apparent predictive power of a theory is the predictive power that this theory provides to Paul regarding the state of his own brain. Here, we are talking about Paul trying to predict the state of his own brain directly "from the inside".

Within the theoretical framework of PP, this new definition of the apparent predictive power takes on its full meaning. Remember that each level of Paul's cortical hierarchy makes predictions about the activity of the level below. Put another way, the brain of Paul constantly seeks to predict its own state. If we assume that Paul is "reducible" to his brain (at least in some sense), this literally means that Paul makes predictions about the state of his own brain.[21] The notion of apparent predictive power refers precisely to this predictive activity. While any cognitive activity is supposed to involve such descending predictions, we suggest that the practice of science by Paul—the testing of scientific hypotheses or theories by Paul—is no different. In short, when Paul confronts a theory with an observation, this activity could literally be described in terms of top-down predictions and bottom-up prediction errors operating in his brain (see Figure 5). To some extent, this means that, the more accurate these top-down predictions are, the greater the apparent predictive power of the theory *from which these predictions derive*. Note that the idea here is to describe the practice of science *per se*, not to draw an analogy (Ma et al., 2023).[22]

An important question arises: In this context, what is a *theory* and what is an *observation*? The foregoing presupposes that a scientific theory is merely a high-level hypothesis operating within the brain of Paul (see Footnote 10). This idea is fairly intuitive. It is a matter of considering that a scientific theory is a special case of—abstract and amodal—hypotheses about the external world. The purpose of such high-level hypotheses is indirectly to explain complex statistical regularities occurring in Paul's sensory inputs. Remember that these statistical regularities stem from causal regularities existing in the external world.

On the other hand, an observation would only be a hypothesis operating at lower levels of Paul's cortical hierarchy (e.g., the representation of blue). More precisely, a low-level hypothesis could be considered as an observation if and only if it has the highest posterior probability (i.e., if and only if it is one of the winning hypotheses). According to our identity hypothesis, these low-level hypotheses are supposed to be (perceptual[23]) *first-person* observations. This raises the question: What would be for Paul a *third-person* observation?

[21] With this view, it seems that the highest level is not subject to any inference, hence the notion of "cognitive core" in Sandved-Smith and Da Costa, 2024; see also Friston et al., 2012.

[22] Describing how an agent practices science could benefit the project of *naturalized epistemology* (Quine, 1969; Rysiew, 2016).

[23] Keep in mind that our identity hypothesis states that the subjective experience of Paul is the set of *all* winning hypotheses occurring within his brain (across all levels). In the following, when referring to first-person observations, we will refer to perceptual experience, and therefore to low-level winning hypotheses, unless mentioned otherwise.

6.1 What is a third-person observation?

To fully appreciate the notion of third person observation, a mere definition is not sufficient. We also need to ask ourselves the following question: When Paul is dealing with third-person observations, does the apparent predictive power of a theory reflect its real predictive power? This question is at the heart of the current subsection.

A third-person observation has something to do with the external world for two reasons. First, by definition, it is about observing something that takes place in the external world. Second, a third-person observation is *in itself* something that takes place in the external world. For instance, think about Paul conducting a neurophysiological measurement. In this context, the *observed thing* would be the brain of another person. On the other hand, what we call third-person observation would be the outcome of the neurophysiological measurement. More precisely, to conduct a neurophysiological measurement, Paul must use a measurement device (e.g., an electroencephalogram). When the measurement device operates, what is happening within the observed brain affects its state. Hence, the state of the measurement device provides information about what is happening within the observed brain. What we refer to as a third-person observation could therefore be defined as the state of the measurement device at the time of measurement. Of course, third-person observations do not always involve a measurement device. Paul can also observe something in a more direct way. In this case, what we call a third-person observation could be defined directly as the state of the observed thing. In any case, whether or not Paul uses a measurement device, a third-person observation is always in itself something that takes place in the external world.

Generally speaking, the real predictive power of a theory regarding third-person observations is the level of fit between third-person observations and the implications of this theory regarding third-person observations. However, this definition needs to be developed further, given that we have defined third-person observations differently depending on whether or not Paul uses a measurement device. On the one hand, when Paul uses a measurement device, the real predictive power of a theory is the level of fit between the state of the measurement device and the implications of this theory regarding the state the measurement device. On the other hand, when Paul does not use a measurement device, the real predictive power of a theory is directly the level of fit between the state of the observed thing and the implications of this theory regarding the state of the observed thing.

Now, let us consider the (more tricky) question of the apparent predictive power. The issue is that Paul's brain does not have direct access to the external world. Paul's access to third-person observations is always mediated by his sensory inputs and subsequently by hypotheses (about the external world) operating at *lower levels* of his cortical hierarchy. Under our identity hypothesis, this literally means that Paul's access to third-person observations is always mediated by first-person observations (see Figure 5). The basic idea here is that the only thing to which Paul has a direct access are first-person observations. Hence, when Paul uses a theory to predict third-person observations, in practice he is only predicting first-person observations (i.e., hypotheses operating at lower levels of his cortical hierarchy)[24]. Let us consider a concrete example. When Paul is expecting

[24]On this subject, some authors would say that Paul is a *naive realist*: that is, he considers first-person observations as being *directly* the external world, in some sense (see Crane and French, 2021). Here, it is not a matter of considering that Paul explicitly defends the theoretical or philosophical position of naive realism. Rather, it is about considering that every human being in their daily life *implicitly* considers their perceptual experience as being directly the external world.

an object in front of him—in the external world—to be blue (understand here blue as an objective property, i.e. of the external world), in practice he is expecting the representation of blue to occur at lower levels of his cortical hierarchy. This leads us to the following statement: There is no fundamental difference between the cognitive activity of predicting third-person observations and the cognitive activity of predicting first-person observations. In both cases, this activity could be described in terms of top-down predictions operating within the brain of Paul. As a consequence, whether we are talking about first-person observations or third-person observations, the apparent predictive power of a theory has something to do with the *accuracy* of these top-down predictions.

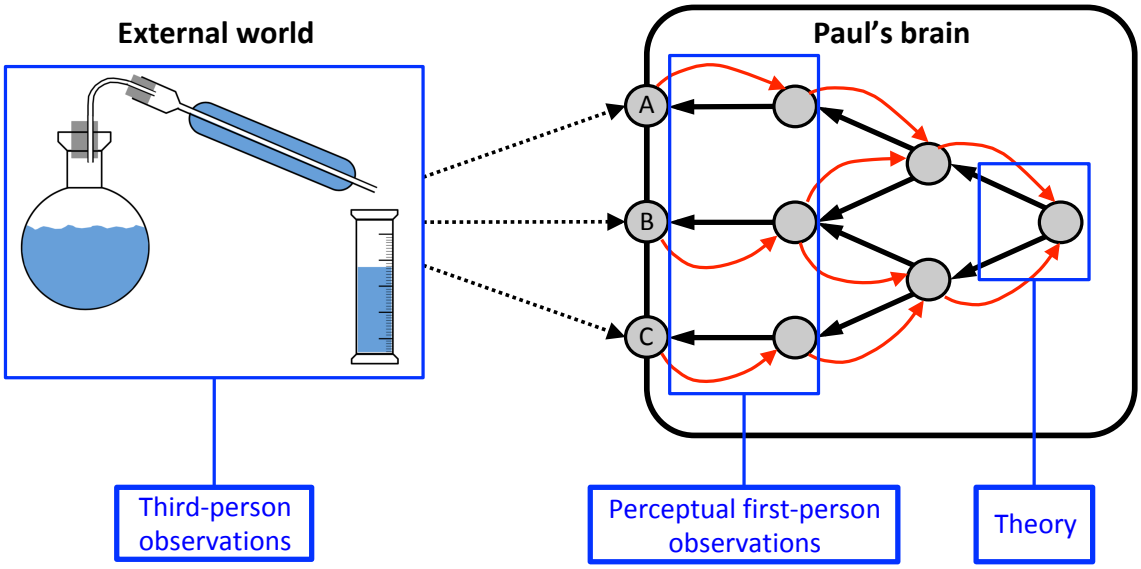

Figure 5. *Very simplified representation (of the brain) of Paul testing a theory against third-person observations.* The circles containing the letters A, B and C correspond to sensory inputs. Each of these inputs belongs to a different sensory modality. The winning hypotheses corresponding to *perceptual* first-person observations are those of the lowest level of Paul's cortical hierarchy (at least in this very simplified representation). Crucially, this representation puts into perspective the fact that Paul's access to third-person observations is always mediated by its sensory inputs and subsequently by perceptual first-person observations.

We are now ready to answer the question: When Paul is dealing with *third-person* observations, does the apparent predictive power of a theory reflect its real predictive power? To answer this, let us consider an example. For the sake of simplicity, this example is not about a *scientific* hypothesis or theory but only about a hypothesis that we are likely to use in everyday life. Paul finds himself in front of a French flag. However, the momentary hypothesis of Paul is that it is an Italian flag (i.e., this incorrect hypothesis has the highest *prior* probability). The key point here is precisely that Paul is using an incorrect hypothesis. As a consequence, this hypothesis does not have a perfect *real* predictive power. More specifically, there is a mismatch between the implications of this hypothesis regarding the leftmost color of the flag (green) and the real leftmost color of this flag (blue).

Crucially, what we are referring to here has nothing to do with Paul's brain. Even if Paul's brain had never existed, we could still talk about the real predictive power of the hypothesis "It is an Italian flag"—one could still say that there is a mismatch between

the implications of this hypothesis and the leftmost color of the flag.

Regardless, if the apparent predictive power of this hypothesis reflects its real predictive power, then this mismatch must result in large bottom-up prediction errors *within the brain of Paul*. In other words, prediction errors regarding the real predictive power of this hypothesis must result in prediction errors regarding its apparent predictive power.

As a matter of fact, in such a situation, large prediction errors would indeed occur in Paul's brain. It is precisely through these bottom-up prediction errors that Paul's higher-level beliefs would be updated, allowing the correct hypothesis (i.e., "It is a French flag") to take over.

In summary, when Paul is dealing with third-person observations, the apparent predictive power of a theory indeed reflects its real predictive power. This simply means that when a theory is able to account for a third-person observation, it gives Paul a certain predictive power.[25] Having said this, it should be noted that third-person observations do not completely escape the epistemological hypothesis for two reasons. First, Paul's access to third-person observations is always mediated by first-person observations. Second, as we will see, when it comes to first-person observations, the epistemological hypothesis is true. In Section 8.2, we discuss the consequences of the fact that third-person observations cannot completely escape the epistemological hypothesis.

Now, let us take a closer look at the bidirectional relationship between high-level hypotheses (e.g., a scientific theory) and low-level hypotheses (i.e., first-person observations). This involves the notion of *causal properties*.

6.2 The notion of causal properties

For the sake of simplicity, in what follows, we assume that the representations of blue and green are two hypotheses of a categorical belief that belongs to the lowest level of Paul's cortical hierarchy.

At any given level of the cortical hierarchy, posterior beliefs are used as priors for the level below and as sensory evidence for the level above. In particular, this means that lower-level beliefs are likely to affect the updating of higher-level beliefs. The representation of blue can therefore be described in two ways. We can either describe it directly, or indirectly through its effects on the updating of Paul's higher-level beliefs: The representation of blue is the hypothesis which, in a certain context, implies that Paul's higher-level beliefs are updated in a certain way. A concrete example might help: Paul expects to see the Italian flag, but the flag finally presented to him is identical to the Italian flag except that the leftmost color is not green but blue. Prediction errors are then generated challenging the high-level hypothesis "It is an Italian flag" in favor of the hypothesis "It is a French flag" (i.e., the hypothesis "It is a French flag" would have the highest *posterior* probability). The key point here is that the manifestation of the representation of blue at lower levels of Paul's cortical hierarchy gave rise to a particular update of his higher-level beliefs. In this regard, we will henceforth talk about the "causal properties" of the representation of blue. By causal properties of blue, we must therefore understand the particular way in which the representation of blue affects the updating of

[25] Here, we ignore all the subtleties that make this statement not always true in practice. Let us consider an example. To deduce the implications of a theory, it is sometimes necessary to perform computer simulations. The issue is that the computational cost of some simulations may be too high and, therefore, impossible to perform. In such cases, Paul could not deduce the implications of the theory. As a consequence, this theory could have been able to account for a third-person observation without providing Paul with any predictive power regarding this observation.

Paul's higher-level beliefs. Crucially, there is another way to describe the causal properties of the representation of blue. Rather than focusing on the influence of this representation on the updating of Paul's higher-level beliefs, we could instead examine the propensity of this representation *to be predicted* by Paul's higher-level beliefs. It amounts to the same. Indeed, in the example we have just used, if the representation of blue gave rise to the high-level hypothesis "It is a French flag", it is precisely because this hypothesis predicts the representation of blue and therefore minimizes prediction errors.

An important question arises: Why does the representation of blue have the causal properties it does? Note that these causal properties are merely the reflection of the generative model parameters: that is, of the synaptic weights that encode causal regularities of the external world. Hence, if the representation of blue has the causal properties it does, it is indirectly because of causal regularities existing in the external world. If the representation of blue gave rise to the hypothesis "It is a French flag", it is because Paul has learned that—in the external world—the leftmost color of the French flag is the color blue.

As we have already mentioned, this learning is indirect. Paul does not have direct access to the blue present in the external world. This learning is mediated by the representation of blue itself and, before that, by Paul's sensory inputs. More precisely, the idea is as follows:

1. Causal regularities existing in the external world give rise to statistical regularities in Paul's sensory inputs.

2. These statistical regularities in turn give rise to other statistical regularities occurring at lower levels of Paul's cortical hierarchy[26]. Importantly, some of these lower-level statistical regularities involve the representation of blue.

3. These lower-level statistical regularities are then learned by the higher levels of Paul's cortical hierarchy. The causal properties of the representation of blue stem directly from this last step.

This leads us to the following statement: The causal properties of the representation of blue stem from the learning by Paul's brain of the *statistical* properties of the representation of blue. Note that, by "statistical properties of the representation of blue", we must understand the set of all statistical regularities occurring at lower levels of Paul's cortical hierarchy and involving the representation of blue.

We are now ready to demonstrate that PP implies the objective epistemological hypothesis.

6.3 A third-person version of the inverted spectrum thought experiment

To determine whether the objective epistemological hypothesis is true in the context of PP, all we have to do is compare the apparent predictive power of a theory with its real predictive power. It is simply a matter of determining whether the former reflects the latter. In fact, we have already applied this procedure in Section 6.1. The only difference is that we are now interested in Paul dealing with *first-person* observations and not third-person ones. Hence, the real predictive power no longer concerns the external world, but

[26] If statistical regularities in sensory inputs give rise to statistical regularities in lower levels, it is because these lower levels track the statistical regularities of sensory inputs.

directly the state of Paul's brain. More specifically, our focus is now on the occurrence of the representations of blue and green at the lowest level of Paul's cortical hierarchy.

Let us imagine that Theory A is a perfect theory of both Paul's brain and the interaction between Paul's brain and the external world. There is therefore a perfect match between the state of Paul's brain and the implications of Theory A regarding the state of Paul's brain. For example, when Theory A predicts that the representation of blue will occur in Paul's brain, the representation of blue indeed occurs in Paul's brain. By definition, this means that Theory A has a perfect real predictive power.

What about the apparent predictive power of Theory A? Does the apparent predictive power of Theory A reflect its real predictive power? This is not an easy question to tackle. In principle, a high-level hypothesis operating in Paul's brain (e.g., Theory A) can very well give rise to accurate top-down predictions regarding the representations of green and blue. However, remember that when we were in Paul's shoes and had a first-person POV (in Section 2), while we began by showing that Theory A could allow us to predict first-person observations (at least in principle), we used in a second step the HPC thought experiments to show that this predictive power was in fact partly based on our prior knowledge of the appearance of blue. Here, we will proceed in a similar way. Let us start by assuming that Theory A (as a high-level hypothesis) enables Paul to obtain *very accurate* top-down predictions about the representations of green and blue. This seems to presuppose that the apparent predictive power of Theory A indeed reflects its real predictive power. In order to challenge this supposition, let us conduct the following thought experiment.

Imagine a world where the posterior probabilities of the representations of green and blue are always inverted. Put another way, imagine a world where the representation of green systematically occurs in Paul's brain instead of the representation of blue (and vice versa). Of course, when we say that the representation of green occurs *instead* of the representation of blue, it is in comparison with the "real" world. The key point here is that this thought experiment describes a world in which Theory A is not true. In constrast to the real world, Theory A does not have a perfect real predictive power. When Theory A predicts that the representation of blue will occur in Paul's brain, it is the representation of green that occurs instead (and vice versa). This does not mean that Theory A is entirely wrong, but that there is at least something it does not account for. For example, perhaps Theory A does not consider the fact that Paul unknowingly wears a device that inverts blue light and green light. Such a device would be sufficient to invert the posterior probabilities of the representations of green and blue, if Paul has been wearing it since the beginning of his life.[27]

Whatever underlies this inversion, let us ask: In this world, what predictive power would Theory A provide to Paul regarding the representations of blue and green? Crucially, if the apparent predictive power of Theory A reflects its real predictive power, then Theory A must lead Paul to make erroneous predictions about these representations. More precisely, the fact that there is a mismatch between observations (i.e. the occurrences of the representations of blue and green) and the implications of Theory A must result in large bottom-up prediction errors within the brain of Paul. In short, prediction errors regarding the real predictive power of Theory A must result in prediction errors

[27] In order to compute a posterior probability using Bayes' theorem, one must use a *likelihood* function mapping the external world to sensory inputs, as well as a *prior* about the external world. Importantly, both the likelihood and the prior have been learned in the inverted world so that they are inverted. This implies indeed an inversion of the posterior probability. See main text for further developments.

regarding its apparent predictive power.

The problem is that, in such an "inverted" world, the statistical properties of the representations of green and blue would be reversed in comparison to the real world. This would be a direct consequence of the fact that the representation of green systematically occurs in Paul's brain instead of the representation of blue (and vice versa). Remember that the *causal* properties of the representations of green and blue stem from the learning of the statistical properties of these representations. As a consequence, the causal properties of the representations of green and blue would also be reversed in comparison to the real world. It is important to understand what this means.

First, the particular way in which the representations of green and blue would affect the updating of Paul's higher-level beliefs would be reversed in comparison to the real world.[28] For example, the representation of blue would no longer constitute an evidence for the hypothesis "It is a French flag" but would be one for the hypothesis "It is an Italian flag".[29] Hence, although Paul wears a device that inverts blue and green light, his ability to recognize a French flag would not be impaired. More generally, when confronted with colored objects, he would behave exactly in the same way as he does in the real world.

Second, the propensity of the representations of green and blue *to be predicted* by Paul's higher-level beliefs would be reversed. For example, the hypothesis "It is an Italian flag" would no longer predict the representation of blue but the representation of green. Of course, this inversion also applies to the top-down predictions resulting from Theory A. Remember that Theory A is nothing more than a high-level hypothesis operating in Paul's brain. It is worth noting that this change does not mean that the Theory A used by Paul in the inverted world is different from the one he used in the real world. If we were to ask Paul to describe the theory he is currently using (i.e., Theory A), no matter whether we were in the real or inverted world, Paul would describe exactly the same theory down to the smallest detail.

In summary, *both* observations (i.e., the occurrences of the representations of green and blue) and the predictions of Paul regarding these observations would be reversed in comparison to the real world. The representations of green and blue could have been "inverted" (violating the "real" predictions of Theory A), this would have made no difference to Paul in terms of prediction errors and belief updating. As in the real world, the top-down predictions resulting from Theory A would *not* have given rise to (large) bottom-up prediction errors. Put another way, Paul would be blind to the fact that his first-person observations contradict the (real) predictions of Theory A.[30]

As we can see, this thought experiment describes a situation in which prediction errors regarding the real predictive power of Theory A does *not* result in prediction errors regarding its apparent predictive power. The conclusion of this thought experiment is therefore that the apparent predictive power of Theory A does not reflect its real predictive

[28] The *same* representation can have *different* causal properties (as defined in Section 6.2) because the generative model encoded by synaptic weights, upon which belief updating depends, is different (see Footnote 27).

[29] Such an inversion of causal properties (between the representations of green and blue) implies that it is when the representation of green occurs in his brain that Paul would tell us "I see the color blue." Under our identity hypothesis, this also means that Paul would call "blue" a color that literally looks like the color we call "green" in the real world.

[30] It should go without saying that these statements only concern the first-person POV (see Section 6.1). From the third-person POV, nothing would prevent us (or Paul) to find that the predictions of Theory A are erroneous. Indeed, by looking at Paul's brain, one could clearly see that the occurrences of the representations of green and blue are inverted compared to the predictions of Theory A.

power. In short, PP implies the objective epistemological hypothesis. The fact that the apparent predictive power of Theory A reflects (or not) its real predictive power rests on the ability (or inability) of Paul to deduce the implications of Theory A, "from the inside". Hence, it could be objected that we have not considered and explored all the possible cognitive processes (e.g., counterfactual and logical reasoning) and external processes (e.g., computer simulation) that Paul could have leveraged to deduce the implications of Theory A. However, this does not compromise the conclusion of our thought experiment for one simple reason: Whatever the way Paul deduces the implications of Theory A, the resulting predictions will always take the form of top-down predictions operating in his brain. In principle, this means that, no matter how Paul deduces the implications of theory A, the ensuing predictions will always depend on the causal properties of the representations of green and blue. This is sufficient for our conclusion to remain true. Finally, note that Fazekas and Jakab (2016) propose a similar thought experiment within another theoretical framework and provide valuable additional information (in particular with regard to what they call *functional unanalyzability* and *linguistic inexpressibility*).

6.4 Updating our interpretation of the HPC thought experiments

Since PP implies the objective epistemological hypothesis, the winning hypothesis theory—PP associated with our identity hypothesis—implies the subjective epistemological hypothesis (see Section 4.3). This means that if *both* PP and our identity hypothesis were true, then the subjective epistemological hypothesis would also be true. In short, the winning hypothesis theory predicts that we are unable to deduce what a theory implies regarding first-person observations. Of course, this statement concerns all theories and therefore the winning hypothesis theory itself: The winning hypothesis theory predicts that we are unable to deduce its own implications regarding first-person observations. This is already an important result. In this context, the fact that the winning hypothesis theory does not allow us to predict first-person observations is no longer a reason to think that this theory is unable to account for these observations. In the context of the winning hypothesis theory, the explanatory hypothesis is no longer a satisfactory interpretation of the HPC thought experiments.

6.5 Absolute properties and relative properties

At this point, we need to clarify a critical detail. Until now, we have always argued that we are unable to deduce what a theory implies regarding first-person observations. In fact, this statement is not entirely true.

Let us consider a concrete example. Under the winning hypothesis theory, our subjective experience of blue is nothing more than the representation of blue. This theory enables us to formulate the following prediction: Each time the representation of blue occurs in our brain, we see (i.e., experience) the color blue. However, as pointed out in Section 2, if we had never seen the color blue in our life, we would not know what it is like to see blue and we could not translate this prediction into an actual expectation about what we see when the representation of blue occurs in our brain. Crucially, this does not mean that the winning hypothesis theory would not give us any information about first-person observations. Indeed, even without knowing what the color blue looks like, we could still make the following prediction: Each time the representation of blue occurs in

our brain, we always see the *same* color. As an example, let us assume that the winning hypothesis theory predicts that the representation of blue occurs in our brain both in situation 1 and in situation 2. Hence, according to the winning hypothesis theory, when we are in situation 1, we are supposed to see the same color as when we are in situation 2. It is important to stress that, even without knowing what it is like to see blue, nothing prevents us to test this prediction directly through first-person observations. It is simply a matter of checking whether or not the color we see in situation 1 is the same as the one we see in situation 2—no matter what this color looks like.

In sum, one could distinguish the *absolute* properties of a particular first-person observation from its *relative* properties. On the one hand, the notion of absolute properties refers directly to *what it is like* to conduct this observation (e.g., what it is like to be in situation 1). On the other hand, the notion of relative properties refers to whether the absolute properties of this observation are identical to (or different from) those of other observations (e.g., the absolute properties of situation 1 are identical to those of situation 2). Be aware that when considering continuous spaces (e.g., color space), the question is rather *how much* the absolute properties of this observation are similar or different from those of other observations. This notion of relative properties is at the heart of the *structuralist methodology*. In particular, this methodology consists in measuring and then modeling the relative differences between subjective experiences (e.g., of colors) using a mathematical structure known as *quality space* (Clark, 1996; Cohen et al., 2015; Kleiner, 2024b; Lee, 2021; Nosofsky, 1992; Roads and Love, 2024; D. Rosenthal, 2010; Tsuchiya and Saigo, 2021). Regardless, the key point is that the epistemological hypothesis would only concern absolute properties but not relative properties.[31] We would be unable to deduce what a theory implies regarding absolute properties, but nothing would prevent us from deducing what this theory implies regarding relative properties. As we will see, this statement can be derived directly from the winning hypothesis theory.

Let us start with the following question: How should we describe the absolute and relative properties of situation 1 when we adopt a third-person perspective? Under our identity hypothesis, the answer is simple:

- The notion of absolute properties refers directly to the fact that, when Paul finds himself in situation 1, the *representation of blue* occurs in his brain.

- The notion of relative properties refers to the fact that the representation occurring in situation 1 is the *same* as the one occurring in situation 2 (for the sake of simplicity, we consider here that situation 2 is the only other situation in which Paul sees the color blue).

Now, let us assume that the following two statements are true:

1 According to Theory A, when Paul finds himself in situation 1, the representation of blue indeed occurs in his brain.

2 According to Theory A, the representation occurring in situation 1 is indeed the same as the one occurring in situation 2.

[31] A similar idea can be found in Davies, 2021a and Davies, 2021b. Indeed, according to this author, the HPC has something to do with absolute quantities but not relative quantities. Furthermore, some researchers have suggested that absolute properties do not exist or, at the very least, can be reduced to relative properties. This is what Kleiner (2024a) calls *ontic phenomenal structural realism*. If the epistemological hypothesis concerns only absolute properties and not relative ones, then this attempt to avoid the HPC by reducing first-person observations to their relative properties is not surprising.

It is crucial to understand what these two statements mean. According to the first, Theory A has a perfect real predictive power regarding the absolute properties of situation 1. According to the second one, Theory A has a perfect real predictive power regarding the relative properties of situation 1.

A question then arises: What would it be if Paul were living in a world where the representation of green systematically occurs instead of the representation of blue (in comparison to the "real" world) and vice versa? With regard to absolute properties, Theory A would no longer have a perfect real predictive power. More precisely, there would be a gap between the absolute properties of situation 1 and the implications of Theory A regarding the absolute properties of situation 1. This is a direct consequence of the fact that the representation of green would occur instead of the representation of blue in situation 1. Crucially, with regard to *relative* properties, Theory A would still have a perfect real predictive power. The reason for this is simple: The representation of green would occur instead of the representation of blue *both* in situation 1 and in situation 2 so that the representation occurring in situation 1 would still be the same as the one occurring in situation 2. The implications of Theory A regarding relative properties would therefore remain correct. The basic idea here is that, when we move from the real world to the inverted world, absolute properties change, while relative properties are perfectly preserved.

What about the apparent predictive power of Theory A? As pointed out in Section 6.3, Theory A would enable Paul to obtain very accurate top-down predictions *both* in the real world and in the inverted world. This leads us to the following statement: The apparent predictive power of theory A is merely a reflection of its real predictive power regarding *relative* properties. Hence, when we say that the apparent predictive power of theory A does not reflect its real predictive power, we are only referring to absolute properties: *the epistemological hypothesis only concerns absolute properties.*[32]

Note that, in order to conduct this reasoning, we have switched from the first-person POV to the third-person POV using our identity hypothesis. The final step of this reasoning, therefore, focuses on returning to the first-person POV using again our identity hypothesis, as shown in Figure 6.

7 Toward simulating a brain facing the hard problem of consciousness

The aim of this section is to show that the winning hypothesis theory makes it possible to formulate predictions about the neural correlates of the HPC. Remember that these predictions would play a key role in solving the problem of consciousness. It is by testing

[32]This reasoning offers a new insight into the notion of apparent predictive power. If the apparent predictive power of a theory only reflects its real predictive power regarding *relative* properties, then any prediction must be interpreted as a prediction about relative properties. This means that, when we expect to see the color blue in situation 1, we are actually expecting to see the *same* color as the one we see in situation 2. If this prediction about the relative properties of situation 1 takes the form of a prediction about the absolute properties of situation 1, it is simply because we have already found ourselves in situation 2 and, therefore, we already know what it is like to see the color of that situation. Without this prior knowledge, we could not have translated our prediction about the relative properties of situation 1 into a concrete expectation about what we will see in situation 1 (as pointed out in the thought experiment of Mary's room in Section 2). In summary, any prediction about the absolute properties of a first-person observation would be in fact a prediction about the relative properties of this observation combined with our prior knowledge about the absolute properties of other observations.

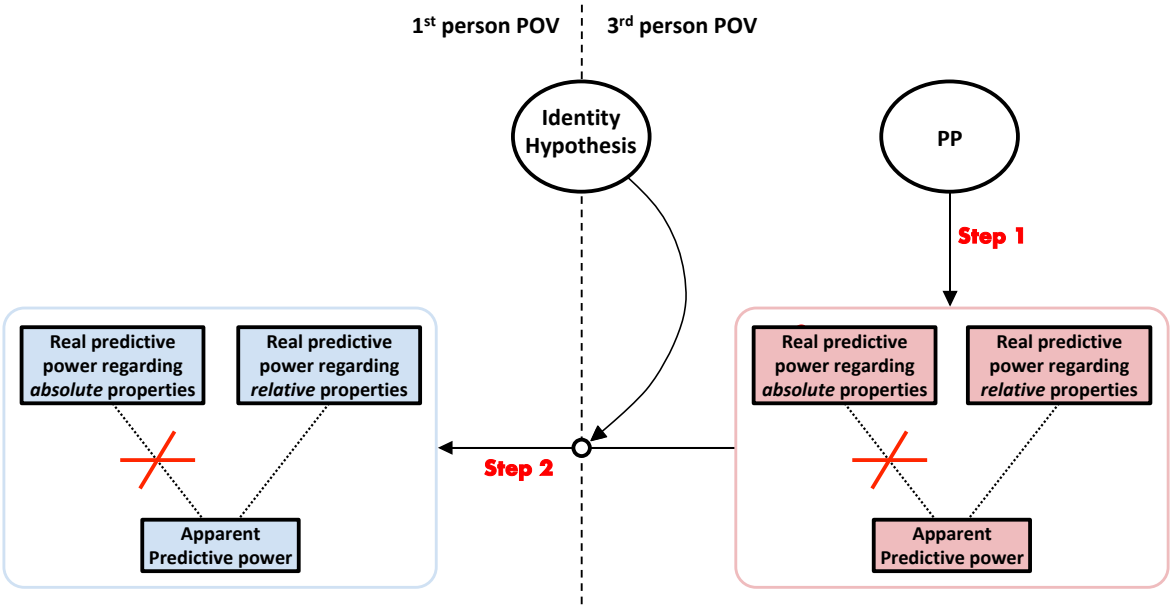

Figure 6. *Two-step reasoning demonstrating that, according to the winning hypothesis theory, the (subjective) epistemological hypothesis only concerns absolute properties.* The vertical dotted line marks the separation between the first-person POV (light blue rectangles) and the third-person POV (light red rectangles). Step 1 consists in showing, from the third-person POV, that PP implies that the apparent predictive power of any theory only reflects the real predictive power regarding *relative* properties, but not the one regarding *absolute* properties. Then, in step 2, using our identity hypothesis, we switch to the first-person POV and conclude that the (subjective) epistemological hypothesis only concerns *absolute* properties.

these predictions that we can assess the validity of our identity hypothesis and therefore the validity of the equivalence between our two definitions of real predictive power (see Sections 4.4 and 4.5). Finally, if our two definitions of real predictive power are indeed equivalent, then we can evaluate the ability of a theory to account for first-person data through evaluating its real predictive power from a third-person POV (which is an "easy" task).

Let us start by remembering the line of reasoning we aim to implement here. The purpose of this reasoning is to deduce what happens in Paul's brain when he conducts the inverted spectrum thought experiment. In principle, all we have to do is answer: Given how it works, how could the brain of Paul implement this thought experiment? The problem is that the thought experiment we are interested in here explicitly refers to first-person observations: "Imagine a world in which the appearance of blue is that of green (and vice versa)". If Paul uses this terminology to describe the "inverted spectrum" scenario, it is because he has a first-person POV, which is not the case for us. The first step in our reasoning is therefore to reformulate this thought experiment using our identity hypothesis (see Section 4.4). The result we are supposed to obtain corresponds precisely to the thought experiment we discussed in Section 6.3. In short, when one takes a third-person perspective, the inverted spectrum thought experiment must be formulated as follows: "Imagine a world in which the representation of green systematically intervenes in Paul's brain instead of the representation of blue (and vice versa)". Crucially, if our identity hypothesis is true, this new version of the inverted spectrum thought experiment

and the original one literally refer to one and the same scenario (i.e., it is the same scenario described from two different POVs). Hence, when Paul conducts the inverted spectrum thought experiment, from our POV we must consider that he is literally conducting the following thought experiment: "Imagine a world in which the representation of green systematically intervened in *my* brain instead of the representation of blue (and vice versa)". One should note that, as we have already mentioned, it is important not to confuse our POV with that of Paul.[33]

Regardless, we can directly use this new version of the inverted spectrum thought experiment to answer our question, namely: *Given how it works, how could the brain of Paul implement this thought experiment?* In what follows, we propose a first approximation of the answer to this question.

Generally speaking, a thought experiment always involves imagining a situation that is not real (Aguinis et al., 2023). For example, Paul could conduct the following thought experiment: What would happen if a car fell from the sky here and now? When Paul performs this thought experiment, there is no car falling from the sky in the external world. This means that a thought experiment consists in a *counterfactual* reasoning. Notably, PP may already be able to account for such counterfactual cognitive activities (Corcoran et al., 2020). In what follows, we will therefore limit ourselves to showing that, if PP is indeed able to account for counterfactual cognitive activities, then it might also have the resources to account for the particular case of the inverted spectrum thought experiment.

When we conduct a thought experiment, we always seek to answer a question of the following form: What would happen in situation X? In fact, it is possible to consider that the *result* of a thought experiment is precisely the answer to this question. The key point is that, unlike the result of a scientific experiment, the result of a thought experiment is not given to us by the external world. On the contrary, we deduce the result ourselves based on our prior knowledge. To take the same example, imagine that we ask Paul: "What would happen if a car fell from the sky here and now?" In this context, Paul would be likely to tell us that this car will continue to fall and then crash into the ground at some point. If Paul is able to formulate this answer, it is because he has learned how the external world works. With PP, this means that, in order to deduce the result of a thought experiment, the brain of Paul relies on previously learned causal regularities (i.e., an agent can leverage its generative model to deduce what would happen in a specific counterfactual situation[34]). In the case of a thought experiment involving the external world, these causal regularities are the ones underlying Paul's sensory inputs. The reason for this is simple: Paul's sensory inputs are the consequences of what is happening *in the external world*. However, this raises the following question: What causal regularities are at play in the inverted spectrum thought experiment? These causal regularities cannot be those underlying Paul's sensory inputs since this thought experiment does not concern the external world but the brain of Paul itself. In the following three sections, we argue that the previously learned causal regularities Paul relies on when performing the inverted spectrum thought experiment are those underlying the amplitude of prediction errors operating within his brain.

[33]Here, it is about Paul conducting this thought experiment *from the first-person POV*. If we describe the thought experiment of Paul without referring to first-person observations, it is no more and no less than because *we* have a third-person POV (no matter the POV of Paul).

[34]For instance, in the context of active inference, when agents compute expected free energies, they anticipate the potential consequences of this or that course of action (Parr and Friston, 2019b).

7.1 The processing fluency theory

As we have already mentioned in Section 5.1, prediction errors are weighted according to their estimated precision. The more precise (i.e., reliable) a prediction error is in a specific context, the larger its weight and the stronger its influence on belief updating.

Crucially, "To improve estimates of precision, precision itself can be predicted; to do so, one must have beliefs about the precision of hierarchically subordinate prediction errors" (Servajean and Wiese, 2024, p. 6; see also Limanowski and Friston, 2018). This is an important observation: The brain of Paul encodes beliefs about the precision of its own prediction errors. Let us turn our attention on how such beliefs would be updated. Imagine that some large prediction errors persist despite their high estimated precision. This situation is likely to arise in particular when the estimated precision is higher than the true precision (i.e., the true reliability of prediction errors). Beliefs about precision can therefore be updated—in a bottom-up way—using the amplitude of prediction errors as evidence. For example, if the amplitude of some prediction errors continues to be large despite their high estimated precision, then Paul can adjust its prior beliefs about precision by forming posterior beliefs about how less precise these prediction errors are.[35]

More generally, it is possible to consider that "having beliefs about precision estimates enables the brain [of Paul] to predict and explain the amplitude of its own prediction errors" (Servajean and Wiese, 2024, p. 7). To some extent, we could speak about hierarchical generative models of prediction errors. Here, it would be more accurate to consider the "rate of prediction error minimization" (i.e., how fast errors are reduced) rather than solely the amplitude of prediction errors (Perrykkad et al., 2021; Van de Cruys, 2017; Van de Cruys and Wagemans, 2011; Van de Cruys et al., 2017, 2022; Velasco and Loev, 2022). However, for the sake of simplicity, we will speak in the following about the amplitude of prediction errors without mentioning the rate of prediction error minimization.

This aspect of PP shares close links with what psychologists call the *processing fluency theory* (Alter and Oppenheimer, 2009). The notion of processing fluency refers to the "level of ease" with which our own cognitive processes occur. According to fluency theorists, our brain constantly infers the causes of processing fluency. This is what is known as the *fluency heuristic*. Many empirical studies have pointed out that the fluency heuristic underpins various cognitive phenomena, such as the feelings of familiarity, valence, and confidence (for an overview, see Alter and Oppenheimer, 2009).

Within the PP framework, processing fluency could be equated with the (inverse) amplitude of prediction errors (Servajean and Wiese, 2024; see also Van de Cruys, 2017; Van de Cruys and Wagemans, 2011; Van de Cruys et al., 2017, 2022; Velasco and Loev, 2022). In this context, the fluency heuristic is an inference about the causes of *prediction errors*. Crucially, this inference is driven by *prediction errors regarding the amplitude of prediction errors* (i.e., the discrepancy between the actual and expected amplitude of prediction errors). The reason for this is twofold.

First, prediction errors regarding the amplitude of prediction errors are used to update beliefs about the causes of prediction errors. Here, fluency theorists would refer to the notion of *relative fluency* and to the *discrepancy-attribution hypothesis* (Whittlesea and Williams, 2000, 2001a, 2001b). Let us consider an example. Imagine that a stimulus gives rise to prediction errors whose amplitude is *lower than expected*. In such a situation, beliefs about the causes of prediction errors would be updated in order to minimize

[35] Beliefs about precision are not always revised downward when prediction errors continue to be large despite their high estimated precision. The reason for this is that beliefs about precision can also be controlled in a top-down way.

this prediction error regarding the amplitude of prediction errors. The hypothesis "I have already encountered this stimulus in the past" (for instance) could then become the hypothesis with the highest posterior probability. Here, the hypothesis "I have already encountered this stimulus in the past" must be understood as a hypothesis about the causes of (low) prediction errors. More precisely, this hypothesis enables Paul to explain *why* the amplitude of prediction errors is so low (see Servajean and Wiese, 2024). The outcome of this process would then be an increase in the estimated precision of prediction errors[36].

Second, prediction errors regarding the amplitude of prediction errors are used to update synaptic weights that encode the causal regularities underlying prediction errors. This is an important observation: The brain of Paul learns the causal regularities underlying its own prediction errors. Put another way, the ability of Paul's brain to explain its own prediction errors stems from a learning process. On this subject, fluency theorists would refer to the notion of *naive theories* (Alter and Oppenheimer, 2009; Briñol et al., 2006; Miele and Molden, 2010; Schwarz, 2004; Thomas and Morwitz, 2009; Unkelbach, 2006).

In what follows, we argue that it is precisely this (learned) knowledge—regarding the causal regularities underlying prediction errors—that Paul would use to deduce the result of the inverted spectrum thought experiment.

7.2 Fluency as an inner sense

It goes without saying that what is happening in the external world affects the amplitude of prediction errors occurring within the brain of Paul. As an example, when Paul comes across a *blue* forest, some large prediction errors persist (i.e., without being resolved) in his brain. The reason for this is that Paul has learned that a forest is not supposed to be blue. However, this should not lead us to conclude that the external world *directly* causes prediction errors. Prediction errors are the result of mismatches occurring within the brain of Paul. Hence, the *direct* cause of prediction errors is not the external world, but rather the brain of Paul "representing" the external world. Considering the same example, the direct cause of persistent prediction errors is not the blue forest, but rather the brain of Paul "representing" the blue forest (i.e., the occurrence of the representation of blue at the lowest level of the cortical hierarchy). This means that the causal relationship between the external world and prediction errors is fully mediated by the brain of Paul. If there are causal regularities between the external world and prediction errors, it is therefore indirectly because there are causal regularities between the brain of Paul—representing the external world—and prediction errors. The key point is that such causal regularities between the brain of Paul and prediction errors can be effective whatever the "real" state of the external world. When the brain of Paul represents a blue forest, some large prediction errors persist, regardless of the real state of the external world (i.e., even if there would not be a blue forest in the external world).

This leads us to the following statement: By learning the causal regularities underlying prediction errors, the brain of Paul learns how it causes itself prediction errors. In this

[36]It should be noted that the prediction errors and precision estimates at sake in this example would not be those of lower perceptual levels (Servajean and Wiese, 2024). Indeed, the precision of lower levels can be affected, for instance, by the visibility of the surroundings but not (or less) by the level of familiarity. As an example, when visual clarity is high, the precision of lower levels is likely to be high even though Paul finds himself in a very unfamiliar environment.

context, beliefs about the causes of prediction errors must be understood as metacognitive beliefs (Lai, 2011) or meta-representations (Kanai et al., 2024). As we will see, in the same way that exteroceptive, interoceptive and proprioceptive senses provide information about the external world (i.e., the body-environment system), allowing the brain to form beliefs and conduct thought experiments about the outside world, fluency would enable the brain to conduct thought experiments about itself, hence the idea of an inner sense. Here, we could also mention the concepts of mental actions and epistemic goals, that is actions and goals about the brain of Paul itself and not the external world (Parr et al., 2023; Pezzulo, 2018).

7.3 The fluency inner sense and the inverted spectrum thought experiment

So far in Section 7, we have put into perspective three important points:

1. To deduce the result of a thought experiment, the brain of Paul relies on causal regularities it has learned.

2. The causal regularities at play in the inverted spectrum thought experiment do not concern the external world but the brain of Paul itself.

3. By learning the causal regularities underlying prediction errors, the brain of Paul learns how it causes itself prediction errors.

This brings us to the following hypothesis: The causal regularities at play in the inverted spectrum thought experiment are those underlying prediction errors. Crucially, if this hypothesis is true, the situation described in this thought experiment should have consequences regarding prediction errors. It is precisely these consequences that Paul is supposed to deduce using his knowledge regarding the causal regularities underlying prediction errors.

In Section 6.3, we have already examined the consequences of this situation with respect to the causal properties of the representations of green and blue. We saw that these causal properties would be reversed in comparison to the real world. Thus, the question is: What does this mean for prediction errors? Remember that the notion of causal properties refers to the particular way in which the representation of blue, for instance, affects the updating of Paul's higher-level beliefs. Crucially, when the representation of blue affects Paul's higher-level beliefs, it does so indirectly through prediction errors. When we talk about the causal properties of the representation of blue, we are therefore implicitly referring to the propensity of this representation to cause this or that prediction error in this or that context. It is important to understand what this means. In the inverted world, the propensity of the representations of green and blue to cause this or that prediction errors would be reversed compared to the real world. In order to make things more intuitive, let us assume that what we usually call "surprise" arises when some large prediction errors persist. This enables us to consider the following example: In the inverted world, Paul would not be surprised to see a blue forest and, conversely, he would be surprised to see a green forest. The propensity of the representation of green and blue to cause (or not) persistent prediction errors would therefore be reversed in comparison to the real world.

In summary, the situation described in the inverted spectrum thought experiment would indeed have consequences regarding prediction errors. It is these consequences

that Paul would deduce using his knowledge about the causal regularities underlying prediction errors. Put another way, the *result* of this thought experiment would be the answer to the question: "In the inverted world, what prediction errors should I expect in this or that situation?" Let us consider a concrete example of the kind of deduction Paul would report: "If I had systematically seen green instead of blue (and vice versa) since the beginning of my life, then I would not be surprised to see a blue forest and, conversely, I would be surprised to see a green forest."

Finally, remember that this thought experiment is supposed to make Paul hypothesize that Theory A is *unable* to account for first-person observations. The key point is that Paul is supposed to put forward this hypothesis even though there is a perfect matching between first-person observations and the implications of Theory A regarding first-person observations (i.e., even though Theory A is able to account for first-person observations). A question then arises: Why does Paul put forward an *incorrect* hypothesis? The answer to this question can be broken down into five steps:

1. The fact that Theory A is able to account for an observation presupposes that if this observation had been different (e.g., "inverted"), then *Theory A would be confronted with prediction errors*. There would be a gap between this observation and the implications of Theory A regarding this observation. Let us call this "Condition X". As mentioned in Section 2.3, a philosopher would formulate Condition X using the notion of *metaphysical impossibility*.

2. Given that Theory A is able to account for first-person observations, it satisfies Condition X. From our POV, this fact is obvious. We can clearly see that, if the representations of green and blue (that is, Paul's first-person observations) had been "inverted", there would be a gap between these observations and the implications of Theory A (see Section 6.3).

3. The problem is that, from Paul's POV, Theory A does not seem to satisfy Condition X. Indeed, as we have seen in Section 6.3, the representations of green and blue could have been "inverted", the top-down predictions resulting from Theory A would *not* have given rise to (large) prediction errors. Here, a philosopher would speak about the notion of *conceivability* (see Section 2.3).

4. By conducting the inverted spectrum thought experiment, Paul would come to believe that Theory A does not satisfy Condition X. More precisely, by deducing what would happen in the inverted world regarding prediction errors, Paul would conclude that Theory A would *not* lead him to make erroneous predictions about the representations of green and blue.

5. Finally, in order to explain why Theory A does not seem to satisfy Condition X, Paul puts forward the hypothesis "Theory A is unable to account for first-person observations". In short, this hypothesis enables Paul to explain why in the inverted world Theory A would *not* have given rise to large bottom-up prediction errors.

As we can see, if Paul puts forward an incorrect hypothesis, it is simply because, from his POV, Theory A does not seem to satisfy Condition X, when in fact it does. Here, a philosopher would say that the reason why Paul comes up with an incorrect hypothesis is the fact that, from his POV, a world in which theory A is true and in which the representations of green and blue are inverted is *conceivable*, whereas such a world is in fact *metaphysically impossible*.

This idea can also be formulated using the distinction between apparent and real predictive power. As we have seen in Section 6.3, the apparent predictive power of Theory A does *not* reflect its real predictive power (at least when it comes to absolute properties). This means that prediction errors occurring in Paul's brain give us information about the apparent predictive power of Theory A *but not about its real predictive power*. As a consequence, the fact that, in the inverted world, Theory A would not have given rise to large bottom-up prediction errors does not tell us anything about the real predictive power of Theory A. So when Paul infers *from this fact* that Theory A is unable to account for first-person observations, he is making a mistake. If Paul puts forward an incorrect hypothesis, it is therefore because he infers the real predictive power of Theory A from prediction errors that do not reflect the real predictive power of Theory A.

To conclude this section, note that the PP framework has a well-established and neurobiologically plausible computational scheme (see the notions of "predictive coding" [Bastos et al., 2012; Friston, Parr, and de Vries, 2017; Rao and Ballard, 1999; Shipp, 2016], "active inference"[37] [Adams et al., 2013; Friston et al., 2009], "planning as inference" [Attias, 2003; Botvinick and Toussaint, 2012; Kaplan and Friston, 2018] and "parametrically deep active inference" [Hesp et al., 2021; Sandved-Smith et al., 2021] as well as "hierarchical models" and "deep temporal models" [Friston, Rosch, et al., 2017[38]]). This means that PP may already have the resources to simulate a brain conducting the inverted spectrum thought experiment. In this context, it is worth noting that there is sometimes a distinction between a computational framework and its actual neurobiological implementation that can provide testable predictions; see the notion of "process theory" in Friston, FitzGerald, et al. (2017). In any case, PP may already allow us to formulate relatively precise predictions regarding the neural correlates of the HPC.

8 Conclusion

Our goal was to answer two questions. First, how can we test experimentally the epistemological hypothesis? Second, under the epistemological hypothesis, how can we solve the HPC? For this purpose, we proposed a method that could allow us to both test experimentally the epistemological hypothesis and solve the HPC (Section 4). We then implemented the first steps of this method in the theoretical framework of PP (sections 5-7). This allowed us to show two important things. First, a theory of consciousness based on PP implies the epistemological hypothesis (Section 6). Second, this theory of consciousness makes it possible to formulate predictions about the neural correlates of the HPC. In particular, in Section 7, we outlined some key elements at play in our brains when facing the HPC. These elements will be further investigated in the future to ultimately

[37] Note that active inference designates either a scheme where actions are selected to maximize model evidence (Friston et al., 2011) or a general theory of *sentient behavior*, encompassing both action and perception (i.e. perceptual inference) (Parr et al., 2022; Pezzulo et al., 2024). Another confusion may also exist between active inference and the free energy principle. In Bayesian mechanics (i.e., the physics of systems encoding probabilistic beliefs) (Ramstead, Sakthivadivel, et al., 2023), the free energy principle shows that, under some conditions depending on the formulation, a system sparsely coupled to another system (i.e., its environment) conforms to active inference (as a theory of sentient behavior) (Friston, Da Costa, Sajid, et al., 2023; Friston, Da Costa, Sakthivadivel, et al., 2023; see also Friston, 2013, 2019). There is also a distinction between active inference as a framework (recipe) and corresponding *process* theories (i.e., descriptions of specific processes implementing the recipe and providing testable predictions) (Friston, FitzGerald, et al., 2017; Pezzulo et al., 2024).

[38] The precise nature of the considered "depth" varies from paper to paper.

simulate a brain facing the HPC and formulate more precise predictions regarding the neural correlates of the HPC. Remember that, if we want to establish the epistemological nature of the HPC and solve this epistemological problem, then the experimental testing of these predictions about the neural correlates of the HPC is a crucial step (refer to Figure 4 for an overview of our method).

As just mentioned, we are aware that these ideas need to be considered in future studies. In this respect, several points deserve to be stressed. First, the implementation of the method we have proposed must be understood only as an *outline* of how we might solve the problem of consciousness. Second, the remarkable compatibility between this method and the theoretical framework of PP does not rule out the possibility that other theories may also be compatible with this method. Third, simulating a brain conducting the inverted spectrum thought experiment may be more difficult to implement than we suggest. Fourth, with regard to the experimental testing of predictions about the neural correlates of the HPC, current neuroimaging techniques may not be sufficient. Fifth, we have not discussed some well-known topics in the neuroscience of consciousness and philosophy of mind (e.g., the notion of *access consciousness* [Block, 1995, 2007]).

Finally, as already pointed out throughout the paper, some of our ideas share close links with existing literature which provides valuable additional information. With this in mind, we discuss next the differences between our ideas and prior related research.

8.1 Links with previous research

Chalmers stated that the problem of consciousness is not reducible to the "easy" problem of consciousness (D. J. Chalmers, 1995, 1996, 2010) (see Footnote 1). If the ideas we have developed are true, then this statement is both true and false. The problem of consciousness would be reducible to the easy problem of consciousness from an explanatory standpoint, but not from an epistemological standpoint. Concretely, this means that a theory that fully solves the easy problem of consciousness is necessarily able to account for our subjective experience, yet this also means that the simple fact of showing that this theory solves the easy problem of consciousness is insufficient to demonstrate that this theory is indeed able to account for our subjective experience. To complete this demonstration, we must also obtain "good reasons" to think that the problem of consciousness is actually reducible to the easy problem of consciousness from an explanatory standpoint. To a certain extent, this is the purpose of the method we have proposed.

As previously discussed, the epistemological hypothesis leads to what philosophers call type-B materialism (or a posteriori physicalism). Some of the new ideas we have developed in Section 4 are probably implicit in at least some versions of type-B materialism, as the phenomenal concept strategy (D. J. Chalmers, 2006; Papineau, 2002; Stoljar, 2005) or the ideas developed in Fazekas and Jakab (2016). Proponents of type-B materialism often argue that, in order to solve the problem of consciousness, one must explain why we say and think that consciousness poses a *hard* problem—one must solve the meta-problem of consciousness. In this paper, we have proposed a solution to the meta-problem of consciousness. However, in our case, this solution (and its experimental testing) corresponds only to a subset of the steps required to solve the problem of consciousness (see Figure 4).

Finally, even if it were proven that our method actually works, this would not mean that we should abandon existing methods used to study consciousness. On the one hand, our method could help us to overcome the hard problem of consciousness (as an epistemological problem) and therefore to provide solid epistemological foundations to the

neuroscience of consciousness. On the other hand, it could help us to clearly identify what is consciousness; that is, to formulate an identity hypothesis and to test experimentally this identity hypothesis. However, even after establishing what our subjective experience is, another important goal is to refine this identity hypothesis, that is to determine what are specifically our subjective experiences of green, chocolate taste and pain (for instance). Under the winning hypothesis theory, this would require to describe precisely the structure of our generative model, and then to determine which hypothesis (or set of hypotheses) corresponds to which subjective experience. To do so, existing methods could be very useful (e.g., computational neurophenomenology [Da Costa et al., 2024; Ramstead et al., 2022; Sandved-Smith et al., 2024]; structuralist methodology [Clark, 1996; Cohen et al., 2015; Lee, 2021; Nosofsky, 1992; Roads and Love, 2024; D. Rosenthal, 2010; Tsuchiya and Saigo, 2021]; see also Seth, 2021).

To conclude this article, let us take a look at one of its most important philosophical perspectives: the subjective foundations project.

8.2 The subjective foundations project

Testing a scientific theory has always been a matter of answering the following question: If this theory were true, would third-person observations be as they are? For instance, if this theory were true, would this measurement device, located in the external world, be in the state it is in? Our entire conception of reality (i.e., the set of all our best theories) relies on this methodology. Crucially, even the reasoning we have proposed in the current paper rests on this methodology as well, as it involves testing predictions about the neural correlates of the HPC.

The issue is that third-person observations are not directly accessible. As we have already mentioned, our access to third-person observations is always mediated by first-person observations (see section 6.1). The basic idea here is that our subjective experience is the only thing we "see". This bring us to the following statement: The only way to justify *rigorously* our conception of reality is to demonstrate that if it were true, then first-person observations—the only thing we "see"—would be indeed as they are. It is fair to consider that this project (i.e., refounding our entire conception of reality on first-person observations) was initiated by Descartes in the 17th century (Descartes, 1885, 1987, 2024; see also the work of Husserl [1970]). Let us call this project the *subjective foundations project*. In what follows, we highlight two important contributions of our paper for the subjective foundations project.

First, it goes without saying that, to complete the subjective foundations project, we need to deduce what our conception of reality implies regarding first-person observations. According to the epistemological hypothesis, we are unable to do so. A question then arises: Does the epistemological hypothesis compromise the subjective foundations project? The answer is no. Remember that the epistemological hypothesis only concerns absolute properties but not relative properties (see Section 6.5). Nothing prevents us, therefore, from deducing—rigorously—what our conception of reality implies regarding the relative properties of first-person observations. Put another way, the epistemological hypothesis does not prevent us from demonstrating that, if our conception of reality were true, then the relative properties of first-person observations would be just as they are. It should be noted that, to have implications about first-person observations, our conception of reality must necessarily include a theory of consciousness (e.g., the winning hypothesis theory).

Having said this, the fact that our conception of reality does not enable us to predict the absolute properties of first-person observations can be problematic. One could conclude from this that our conception of reality is unable to *account* for these absolute properties (which would mean that it is false or incomplete). This problem can be overcome as follows. As we have seen, the winning hypothesis theory implies the epistemological hypothesis (Section 6). Thus, if our conception of reality includes the winning hypothesis theory, then the fact that our conception of reality does not enable us to predict the absolute properties of first-person observations is exactly what we should expect if our conception of reality were true, so that one should not conclude that it is false or incomplete.

These ideas will be further elaborated in future work.

Conflicts of interest

The authors declare no conflicts of interest.

References

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, *218*, 611–643.

Aguinis, H., Beltran, J. R., Archibold, E. E., Jean, E. L., & Rice, D. B. (2023). Thought experiments: Review and recommendations. *Journal of Organizational Behavior*, *44*(3), 544–560.

Ainley, V., Apps, M. A. J., Fotopoulou, A., & Tsakiris, M. (2016). 'bodily precision': A predictive coding account of individual differences in interoceptive accuracy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160003.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and social psychology review*, *13*(3), 219–235.

Attias, H. (2003). Planning by probabilistic inference. *International workshop on artificial intelligence and statistics*, 9–16.

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

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45–53.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1–23.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Battaglia, S., Servajean, P., & Friston, K. J. (2025). The paradox of the self-studying brain. *Physics of Life Reviews*.

Block, N. J. (1982). Functionalism. In *Studies in logic and the foundations of mathematics* (pp. 519–539, Vol. 104). Elsevier.

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

Block, N. J. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and brain sciences*, *30*(5-6), 481–499.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in cognitive sciences*, *16*(10), 485–488.

Briñol, P., Petty, R. G., & Tormala, Z. L. (2006). The malleable meaning of subjective ease. *Psychological Science*, *17*(3), 200–206. https://doi.org/10.1111/j.1467-9280.2006.01686.x

Broday-Dvir, R., Norman, Y., Harel, M., Mehta, A. D., & Malach, R. (2023). Perceptual stability reflected in neuronal pattern similarities in human visual cortex. *Cell reports*, *42*(6).

Byrne, A. (2004). Inverted qualia. *Stanford Encyclopedia of Philosophy*.

Carruthers, P. (2017). Higher-order theories of consciousness. *The Blackwell companion to consciousness*, 288–297.

Cencini, M., Puglisi, A., Vergni, D., & Vulpiani, A. (2021). Prediction. In *A random walk in physics: Beyond black holes and time-travels* (pp. 155–160). Springer International Publishing. https://doi.org/10.1007/978-3-030-72531-0_23

Chalmers, A. (2013). *What is this thing called science?* McGraw-Hill Education (UK).

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

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

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

Chalmers, D. J. (2002). Does conceivability entail possibility? *Conceivability and possibility*, *145*, 200.

Chalmers, D. J. (2003). Consciousness and its place in nature. *The Blackwell guide to philosophy of mind*, 102–142.

Chalmers, D. J. (2006). Nine phenomenal concepts and the explanatory gap. *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism*, 167.

Chalmers, D. J. (2010, October). *The character of consciousness*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195311105.001.0001

Chalmers, D. J. (2018). The meta-problem of consciousness. *Journal of consciousness studies*, *25*(9-10), 6–61.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Clark, A., & Chalmers, D. J. (1998). The extended mind.

Clark, A. (1996). *Sensory qualities*. Oxford University Press.

Cohen, M. A., Nakayama, K., Konkle, T., Stantić, M., & Alvarez, G. A. (2015). Visual awareness is limited by the representational architecture of the visual system. *Journal of Cognitive Neuroscience*, *27*(11), 2240–2252.

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 32.

Crane, T., & French, C. (2021). The problem of perception. *The Stanford encyclopedia of philosophy*.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. J. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, *99*, 102447.

Da Costa, L., Sandved-Smith, L., Friston, K. J., Ramstead, M. J. D., & Seth, A. K. (2024). A mathematical perspective on neurophenomenology. *arXiv preprint arXiv:2409.20318*.

Davies, P. (2021a). Constructing the objective world from subjective perceptions.

Davies, P. (2021b). Why the hard problem of consciousness will never be solved.

Dehaene, S. (2014). *Le code de la conscience*. Odile Jacob.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Descartes, R. (1885). *Les principes de la philosophie: Première partie*. Ch. Delegrave.

Descartes, R. (1987). *Discours de la méthode*. Vrin.

Descartes, R. (2024). *Méditations métaphysiques*. Flammarion.

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why iit and other causal structure theories cannot explain consciousness. *Consciousness and cognition*, *72*, 49–59.

Fazekas, P., & Jakab, Z. (2016). The sensory basis of the epistemic gap: An alternative to phenomenal concepts. *Philosophical Studies*, *173*, 2105–2124.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, *4*, 215.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1–47.

FitzGerald, T. H. B., Dolan, R. J., & Friston, K. J. (2015). Dopamine, reward learning, and active inference. *Frontiers in computational neuroscience*, *9*, 166836.

Fleming, S. M., & Shea, N. (2024). Quality space computations for consciousness. *Trends in Cognitive Sciences*.

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

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*, 25–50.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.

Friston, K. J. (2019). A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*.

Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29.

Friston, K. J., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M. J. D., & Parr, T. (2023). Path integrals, particular kinds, and strange things. *Physics of Life Reviews*.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS one*, *4*(7), e6421.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural computation*, *29*(1), 1–49.

Friston, K. J., & Frith, C. (2015). A duet for one. *Consciousness and cognition*, *36*, 390–405.

Friston, K. J., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological cybernetics*, *104*, 137–160.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network neuroscience*, *1*(4), 381–414.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, *77*, 388–402.

Friston, K. J., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, *3*, 130.

Hardin, C. L. (1988). *Color for philosophers: Unweaving the rainbow*. Hackett Publishing.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural computation*, *33*(2), 398–446.

Hilbert, D. R., & Kalderon, M. E. (2000). Color and the inverted spectrum. *Color perception: Philosophical, psychological, artistic, and computational perspectives*, *9*, 187–214.

Hill, C. (2016). Conceivability and possibility. *The Oxford handbook of philosophical methodology*, 326–347.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, *35*(2), 209–223.

Hohwy, J., Roepstorff, A., & Friston, K. J. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701.

Husserl, E. (1970). *The crisis of european sciences and transcendental phenomenology: An introduction to phenomenological philosophy*. Northwestern University Press.

Hutchinson, J. B., & Barrett, L. F. (2019). The power of predictions: An emerging paradigm for psychological research. *Current directions in psychological science*, *28*(3), 280–291.

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

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, *32*(127), 127–36.

Jackson, F. (1986). What mary didn't know. *The journal of philosophy*, *83*(5), 291–295.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS computational biology*, *9*(6), e1003094.

Kammerer, F. (2021). The illusion of conscious experience. *Synthese*, *198*(1), 845–866.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169.

Kanai, R., Takatsuki, R., Fujisawa, I., & Kanai, R. (2024). Meta-representations as representations of processes.

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological cybernetics*, *112*(4), 323–343.

Kirk, R. (2003). Zombies.

Kleiner, J. (2024a). Topics in mathematical consciousness science.

Kleiner, J. (2024b). Towards a structural turn in consciousness science. *Consciousness and cognition*, *119*, 103653.

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1), niab001.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature reviews neuroscience*, *17*(5), 307–321.

Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.
Lai, E. R. (2011). Metacognition: A literature review. *Always learning: Pearson research report*, *24*, 1–40.
Lee, A. Y. (2021). Modeling mental qualities. *Philosophical Review*, *130*(2), 263–298.
Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific philosophical quarterly*, *64*(4), 354–361.
Lewis, C. I. (1929). *Mind and the world-order: Outline of a theory of knowledge*. Charles Scribner's Sons.
Limanowski, J., & Friston, K. J. (2018). 'seeing the dark': Grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in psychology*, *9*, 643.
Locke, J. (1847). *An essay concerning human understanding*. Kay & Troutman.
Ma, W. J., Kording, K. P., & Goldreich, D. (2023). *Bayesian models of perception and action: An introduction*. MIT press.
Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of consciousness*, *2021*(2), niab028.
Marchi, F., & Hohwy, J. (2022). The intermediate scope of consciousness in the predictive mind. *Erkenntnis*, *87*(2), 891–912.
Maunsell, J. H., & van Essen, D. C. (1983). The connections of the middle temporal visual area (mt) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience*, *3*(12), 2563–2586.
McGinn, C. (1989). Can we solve the mind–body problem? *Mind*, *98*(391), 349–366.
Mesulam, M.-M. (1998). From sensation to cognition. *Brain: a journal of neurology*, *121*(6), 1013–1052.
Miele, D. B., & Molden, D. C. (2010). Naive theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General*, *139*(3), 535.
Miller Tate, A. J. (2021). A predictive processing theory of motivation. *Synthese*, *198*(5), 4493–4521.
Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.
Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, *43*(1), 25–53.
Palmer, S. E. (1999). Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences*, *22*(6), 923–943.
Papineau, D. (2002). *Thinking about consciousness*. Clarendon Press.
Parr, T., & Friston, K. J. (2018). The anatomy of inference: Generative models and brain structure. *Frontiers in computational neuroscience*, *12*, 90.
Parr, T., & Friston, K. J. (2019a). Attention or salience? *Current opinion in psychology*, *29*, 1–5.
Parr, T., & Friston, K. J. (2019b). Generalised free energy and active inference. *Biological cybernetics*, *113*(5), 495–513.
Parr, T., Holmes, E., Friston, K. J., & Pezzulo, G. (2023). Cognitive effort and active inference. *Neuropsychologia*, *184*, 108562.
Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.
Perrykkad, K., Lawson, R. P., Jamadar, S., & Hohwy, J. (2021). The effect of uncertainty on prediction error in the action perception loop. *Cognition*, *210*, 104598.

Pezzulo, G. (2018). Commentary: The problem of mental action: Predictive control without sensory sheets. *Frontiers in Psychology*, *9*, 1291.

Pezzulo, G., Parr, T., & Friston, K. J. (2024). Active inference as a theory of sentient behavior. *Biological Psychology*, 108741.

Prinz, J. (2017). The intermediate level theory of consciousness. *The Blackwell companion to consciousness*, 257–271.

Quine, W. V. O. (1969). Epistemology naturalized. *Knowledge and inquiry*, 245–260.

Ramstead, M. J. D., Albarracin, M., Kiefer, A., Klein, B., Fields, C., Friston, K. J., & Safron, A. (2023). The inner screen model of consciousness: Applying the free energy principle directly to the study of conscious experience. *arXiv preprint arXiv:2305.02205*.

Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), 20220029.

Ramstead, M. J. D., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G., Smith, R., Dumas, G., Lutz, A., Friston, K. J., & Constant, A. (2022). From generative models to generative passages: A computational approach to (neuro) phenomenology. *Review of Philosophy and Psychology*, *13*(4), 829–857.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Ridderinkhof, K. R. (2017). Emotion in action: A predictive processing perspective and theoretical synthesis. *Emotion Review*, *9*(4), 319–325.

Roads, B. D., & Love, B. C. (2024). Modeling similarity and psychological space. *Annual Review of Psychology*, *75*(1), 215–240.

Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues*, *20*, 368–393.

Rosenthal, I. A., Singh, S. R., Hermann, K. L., Pantazis, D., & Conway, B. R. (2021). Color space geometry uncovered with magnetoencephalography. *Current Biology*, *31*(3), 515–526.

Rutar, D., Colizoli, O., Selen, L., Spieß, L., Kwisthout, J., & Hunnius, S. (2023). Differentiating between bayesian parameter learning and structure learning based on behavioural and pupil measures. *PloS one*, *18*(2), e0270619.

Rutar, D., de Wolff, E., van Rooij, I., & Kwisthout, J. (2022). Structure learning in predictive processing needs revision. *Computational Brain & Behavior*, *5*(2), 234–243.

Rysiew, P. (2016). Naturalism in epistemology.

Sandved-Smith, L., & Da Costa, L. (2024). Metacognitive particles, mental action and the sense of agency. *arXiv preprint arXiv:2405.12941*.

Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K. J., Lutz, A., & Ramstead, M. J. D. (2021). Towards a computational phenomenology of mental action: Modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of consciousness*, *2021*(1), niab018.

Sandved-Smith, L., Hohwy, J., Kiverstein, J., & Lutz, A. (2024). Deep computational neurophenomenology: A methodological framework for investigating the how of experience. *Available on https://www. researchhub. com/paper/6171544/deep-computational-neurophenomenology-a-methodological-frameworkfor-investigating-the-how-of-experience*.

Schlicht, T., & Dolega, K. (2021). You can't always get what you want: Predictive processing and consciousness. *Philosophy and the mind sciences*, *2*.

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*(4), 332–348. https://doi.org/10.1207/s15327663jcp1404_2

Servajean, P., & Wiese, W. (2024). Processing fluency and predictive processing: How the predictive mind becomes aware of its cognitive limitations. *Topics in Cognitive Science*.

Seth, A. K. (2021). *Being you: A new science of consciousness*. Penguin.

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452.

Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual review of neuroscience*, *33*(1), 89–108.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in psychology*, *7*, 1792.

Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, *79*(7), 357–381.

Smart, J. J. C. (2000). The mind/brain identity theory.

Smith, R., Schwartenbeck, P., Parr, T., & Friston, K. J. (2020). An active inference approach to modeling structure learning: Concept learning as an example case. *Frontiers in computational neuroscience*, *14*, 41.

Speaks, J. (2018). The space of materialist views.

Stoljar, D. (2005). Physicalism and phenomenal concepts. *Mind & language*, *20*(5), 469–494.

Thomas, M., & Morwitz, V. G. (2009). The ease-of-computation effect: The interplay of metacognitive experiences and naive theories in judgments of price differences. *Journal of Marketing Research*, *46*(1), 81–91. https://doi.org/10.1509/jmkr.46.1.81

Tong, F. (2003). Primary visual cortex and visual awareness. *Nature reviews neuroscience*, *4*(3), 219–229.

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., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), niab034.

Unkelbach, C. (2006). The learned interpretation of cognitive fluency. *Psychological Science*, *17*(4), 339–345. https://doi.org/10.1111/j.1467-9280.2006.01708.x

Van de Cruys, S. (2017). Affective value in the predictive mind.

Van de Cruys, S., Bervoets, J., & Moors, A. (2022). Preferences need inferences: Learning, valuation, and curiosity in aesthetic experience. In *The routledge international handbook of neuroaesthetics* (pp. 475–506). Routledge.

Van de Cruys, S., Chamberlain, R., & Wagemans, J. (2017). Tuning in to art: A predictive processing account of negative emotion in art (commentary). *Behavioral and Brain Sciences*, *40*.

Van de Cruys, S., & Wagemans, J. (2011). Putting reward in art: A tentative prediction error account of visual art. *i-Perception*, *2*(9), 1035–1062.

van Heusden, E., Harris, A. M., Garrido, M. I., & Hogendoorn, H. (2019). Predictive coding of visual motion in both monocular and binocular human visual processing. *Journal of vision*, *19*(1), 3–3.

Velasco, P. F., & Loev, S. (2022). Cognitive feelings in the predictive mind: Emotion, meta-cognition and predictive processing.

Vishne, G., Gerber, E. M., Knight, R. T., & Deouell, L. Y. (2023). Distinct ventral stream and prefrontal cortex representational dynamics during sustained conscious visual perception. *Cell reports*, *42*(7).

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, *1464*(1), 242–268.

Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 547.

Whittlesea, B. W. A., & Williams, L. D. (2001a). The discrepancy-attribution hypothesis: I. the heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 3.

Whittlesea, B. W. A., & Williams, L. D. (2001b). The discrepancy-attribution hypothesis: Ii. expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 14.

Whyte, C. J., Corcoran, A. W., Robinson, J., Smith, R., Moran, R. J., Parr, T., Friston, K. J., Seth, A. K., & Jakob, H. (2024). On the minimal theory of consciousness implicit in active inference. *arXiv preprint arXiv:2410.06633*.

Wiese, W., & Metzinger, T. (2017). Vanilla pp for philosophers: A primer on predictive processing.

Wilkinson, S., Deane, G., Nave, K., & Clark, A. (2019). Getting warmer: Predictive processing and the nature of emotion. *The value of emotions for knowledge*, 101–119.