This is a draft – please do not cite. The final version of this paper is forthcoming in *Philosophical Psychology*

François Kammerer

Can you believe it?
Illusionism about consciousness and the illusion meta-problem

Acknowledgments: I would like to thank Samuel Webb and Joseph Levine for their comments and their help.

Introduction

Phenomenal consciousness is at odds with physicalism. It seems to have many features that we are unable to locate in the physical world, such as qualitativeness or subjectivity. Because of these features of phenomenal states, the "explanatory gap" (Levine, 1983, 2001) and the "hard problem of consciousness" (Chalmers, 1995, 1996) arise, and we find ourselves wondering: how could anything purely physical be phenomenally conscious, or give rise to phenomenal consciousness? Most of us admit that we have no clue as to how to answer this question.

Some people try to solve this problem in a radical way, by saying that phenomenal consciousness is an illusion. In this kind of view, called *illusionism* (Frankish, 2016), phenomenal consciousness merely seems to exist, but does not really exist. All there exists is the physical world, of which the brain is a part, gifted with a peculiar functional organization. But this brain is not genuinely *phenomenally conscious* – none of its states has the qualitative nature or the subjectivity that are so difficult to explain in a physicalist framework. If this is true, then there is no need to explain the emergence of phenomenal consciousness, as phenomenal consciousness is not real and *does not emerge*. All there is to explain is why, given the functional organization of our purely physical mind, it *seems to us* that we are phenomenally conscious even though we are not. In Frankish's terms: the "hard problem of consciousness" is replaced with the "illusion problem" (Frankish, 2016, p. 24).

However, I think that one particular aspect of the illusion problem has been neglected until now. A satisfying illusionist theory of consciousness must not only explain why we have

the illusion that we are conscious, but also why we have *this particular illusion*; notably, why this illusion is so strong, and why it is so difficult for us to recognize its illusory nature. In other words, a satisfying illusionist theory of consciousness has to explain why illusionism itself is so difficult to accept. I will say that a theory that explains this fact solves the *illusion meta-problem*, which I see as an aspect of the more general *illusion problem*. I have already presented this illusion meta-problem in a recent short response paper to Frankish's article on illusionism, in which I called it the "hardest aspect of the illusion problem" (Kammerer, 2016). However, most of this paper was actually devoted to the description of a hypothesis concerning introspection which may allow us to solve this problem, rather than to an argumentation designed to show that this problem is genuine. My goal here is not to solve the illusion meta-problem; it is rather to argue carefully for the fact that this problem is a genuine problem for current illusionist theories of consciousness. I want to show that currently available illusionist theories of consciousness neglect this problem, and fail to solve it: they are unable to explain why illusionism is so difficult to accept. In order to make my case, I will focus on what I take to be two of the most promising currently available illusionist theories of consciousness (Graziano, 2013; Pereboom, 2011), and I will argue that they cannot solve the illusion meta-problem. I don't intend to argue against illusionism in general, but only against its current versions. Indeed, I think that it is possible to solve this problem – I previously suggested a way to do so. However, I think that it is first crucial that we recognize that this problem is real, and that it cries out for a solution. The aim of this paper is therefore to encourage proponents of illusionism to pay more attention to the illusion meta-problem, which I take to present a crucial difficulty for this view.

In a first section, I will present illusionism as a theory of consciousness, drawing mostly on Frankish's recent paper. In a second section, I will focus on two illusionist theories of consciousness (Pereboom's and Graziano's), and I will detail how these two theories try to solve the illusion problem – that is to say, how they explain why it seems to us that we are phenomenally conscious even though we are not. A third section will be devoted to presenting a neglected aspect of the illusion problem: the "illusion meta-problem". In a fourth section, I will argue that Graziano's account cannot solve the illusion meta-problem, and in a fifth section I will argue the same about Pereboom's view. A final section will be devoted to concluding remarks.

1. Illusionism about consciousness

Phenomenally conscious states (or "conscious experiences") are mental states such that there is "something it is like" to be in those states. A visual sensation of red, a gustative sensation of a pear, a feeling of pain in the head, are typical examples of phenomenally conscious states. These phenomenally conscious states are said to possess "phenomenal properties" or "qualia", which are the properties of these states which characterize what it is like for a subject to be in such states. A given mental state is phenomenally conscious if and only if it has phenomenal properties; its phenomenal properties determine what it's like to be in this state.

Many philosophers admit that phenomenal consciousness is at odds with physicalism. Amongst other things, phenomenal properties are often understood as *ineffable and irreducible qualities*; conscious experiences are said to have an *intrinsically subjective mode of existence*, and to be *directly and essentially apprehended* by the subject who has them (Frankish, 2016, p. 13). For all these reasons, it is very difficult to understand how something purely physical could be conscious: how can a purely physical brain, for example, give rise to such a peculiar thing as a *conscious experience*? This rather old problem has been recently labeled the "hard problem of consciousness" (Chalmers, 1995).

Three main strategies are available to philosophers facing this problem (Frankish, 2016, p. 13‑14).[1] First, they can decide to treat consciousness as a real and unique feature of the world, irreducible to anything purely physical (at least in a restrictive understanding of the physical). This "radical realist" stance has many versions, including dualism and neutral monism (which often leads to a form of panpsychism). Second, they can try, in spite of the difficulties, to explain consciousness and its properties in physical terms, using the framework of the empirical science of the mind (such as cognitive neuroscience): this is the "conservative realist" stance. Finally, they can take a third route, called "illusionism".[2] Illusionists deny the reality of phenomenal consciousness, in the sense that they deny that any of our mental states really have "genuine" phenomenal properties. For illusionists about consciousness, strictly speaking, there is "nothing it is like" for any creature to be in any mental state.

Illusionists can still accept that the mental states we call "conscious states" have some properties, "quasi-phenomenal properties" (Frankish, 2016, p. 15), which are purely

[1] In this part of the paper, I will mostly draw on Frankish's paper, and I will use his terminology.

[2] For more details about illusionism, see (Frankish, 2016, p. 14). Here I call "illusionism" what Frankish calls "strong illusionism" – which seems to me to be the most interesting version of illusionism.

physical/functional properties of brain states; they are reliably *tracked* by our introspective devices, but are *mischaracterized* as phenomenal by our introspective representations. In that sense, for any given phenomenal judgment (such as "I am in pain now"), illusionists can give an interpretation of this judgment such that the judgment will often be true (provided that the concept used, here <pain>, is interpreted as bearing on a quasi-phenomenal state, which has a quasi-phenomenal property). [3] However, illusionists maintain that no one is phenomenally conscious strictly speaking, and that strictly speaking all phenomenal judgments are *false* in the actual world.[4]

For illusionists, none of our mental states instantiate phenomenal properties, and there is nothing it is like for us to be in any of our mental states. For that reason, phenomenal consciousness does not need to be explained, as it does not exist. Only quasi-phenomenal consciousness exists, and quasi-phenomenal consciousness is not problematic for physicalism. All that still needs to be explained is why it seems to us that we are phenomenally conscious, while we are not. As Frankish puts it: "Illusionism replaces the hard problem with the illusion problem – the problem of explaining how the illusion of phenomenality arises and why it is so powerful" (Frankish, 2016, p. 37).This means that illusionists have to provide a theory which explains why we tend to judge that we are phenomenally conscious. This theory can appeal to built-in, hard-wired features of our introspective mechanisms (Graziano, 2013; Humphrey, 2011; Pereboom, 2011), to a kind of mistaken inferential mechanism of projection (Rey, 1995), or to a mix of features of our introspective mechanisms and of philosophical (mostly Cartesian) prejudices (Dennett, 1988, 1991).

In this paper, I will only focus on theories that assert that the reason why it falsely seems to us that we are phenomenally conscious is to be found in some hard-wired features of our introspective mechanisms. Indeed, I think that this kind of theory is in a better position to account for the robustness of the illusion that we are conscious: they are more likely to solve the illusion problem. However, I think that the criticisms I will later address to these theories could also easily apply to other forms of illusionism.[5]

---

[3] For this reason, illusionists are not technically committed to eliminativism (Frankish, 2016, p. 21‑22). See also (Pereboom, 2011, p. 43‑46) for more on this "reinterpretation" of concepts of phenomenal properties – understood by Pereboom as a distinction between the "edenic" content and the "ordinary" content of phenomenal concepts.

[4] By "strictly speaking" I mean: if we do not re-interpret the terms used, so that they refer to quasi-phenomenal states.

[5] Ideally, these two claims ((a) theories that explain the illusion of consciousness by appealing to some hard-wired features of our introspective mechanisms are better than the others, and (b) the criticisms I will address to these theories could also apply to other forms of illusionism) should be argued for, but such an argument would be beyond the scope of this paper. However, I think that the point I am making in this paper can be of interest

2. Explanations of the illusion of consciousness

I think that the theories that state that the illusion of consciousness is to be explained by some hard-wired features of our introspective mechanisms are in a better position to solve the illusion problem. However, I think that even these theories – at least in their current versions – are still unable to fully solve this problem. In this section, I will present what I take to be the two most promising illusionist theories of consciousness, and I will quickly describe how they are supposed to solve the illusion problem. In the next section, I will detail the aspect of the problem that these theories don't solve.

The first of the two theories is Pereboom's theory, which he labels the "qualitative inaccuracy hypothesis" (Pereboom, 2009, 2011). Pereboom does not explicitly endorse this hypothesis, even though he tries to make the case that it constitutes an open possibility – for reasons of simplicity though, I will speak of this view as if Pereboom endorsed it. According to this hypothesis, our introspective mechanisms systematically misrepresent phenomenal states. They represent them as having phenomenal properties, gifted with a qualitative nature that they lack in reality. In Pereboom's understanding of the hypothesis, phenomenal properties *really are instantiated*, but they are misrepresented by introspection as having a qualitative nature that they don't have. However, it is possible to slightly reinterpret this hypothesis and to say that these *really instantiated* phenomenal properties (devoid of the qualitative nature that introspection presents them as having) really are simply "quasi-phenomenal properties", so that Pereboom's view fits more nicely with the framework I have discussed (which mostly draws on Frankish's work). On this interpretation, Pereboom's theory has the consequence that *genuine* phenomenal properties, that is to say, properties endowed with the qualitative nature that our introspective mechanisms represent, are never instantiated. However, it *seems to us* that they are instantiated; and it will still seem to us that they are instantiated if we come to *believe* Pereboom's hypothesis. Indeed, our introspective mechanisms are hard-wired in such a way that they will keep on representing us as being in these states, so that we will still be enticed to judge that we are.

The other major illusionist theory I want to discuss is Graziano's theory (Graziano, 2013), which he labels the "attention schema theory". According to this view, our brain is able to monitor its own attentional processes, and to form a schematic representation of these

even if one rejects these two claims. Indeed, in that case, one can simply understand my point as a point about some illusionist theories of consciousness (which happen to be amongst the most popular).

processes, called the "attention schema". This representation is schematic and simplified: instead of representing attentional processes in all their complexities, it represents a simple relation of "awareness" between a subject and a piece of information. This relation is represented as consisting in "a fluidic substance, […] an experience, […] a sentience" (Graziano, 2013, p. 80). But this representation inaccurately depicts our actual attentional processes: our attention schema is *fallacious*, as there is nothing like this simple relation of awareness in our brain, but only a set of complex attentional processes by way of which a cognitive system takes on and processes information concerning some objects. This attention schema constitutes a kind of introspective and permanent representation of our own conscious states: because we have this attention schema, it seems to us that we are in such simple relations of awareness with pieces of information, relations which consist in a fluidic substance, an experience, a sentience. What we call "phenomenal states" are mental states of being in these relations represented by our attention schema. But, as there is no such simple relation of awareness taking place in our brains, there are no such things as phenomenal states. As Graziano later bluntly put it: "Consciousness doesn't happen. It's a mistaken construct" (Graziano, 2016).

Pereboom's and Graziano's views roughly rely on the same general idea.[6] They all say that our introspective mechanisms, because of some of their hard-wired features, represent us as being in mental states endowed with some special properties – properties that they don't really have. These properties happen to be what we call "phenomenal properties". So, for this reason, it systematically seems to us that we are phenomenally conscious, even though we are not.

These theories claim to solve the "illusion problem", by appealing to hard-wired features of our introspective mechanisms. However, I want to show that they fail in this task, because they don't solve one important aspect of this illusion problem. I will call this aspect of the illusion problem, which I have already briefly presented elsewhere (calling it "the hardest aspect of the illusion problem (Kammerer, 2016, pp. 125-126), the "illusion meta-problem". The illusion meta-problem is the problem of explaining some peculiar aspects of the way in which it falsely seems to us that we are conscious (the *mode* of the illusion), namely, the fact that illusionism itself regarding consciousness seems so radically implausible, deeply puzzling, and almost absurd to many. In the following section, I will

[6] And this would be the case of other theories too (Humphrey, 2011).

explain in details what the illusion meta-problem is. I will later argue that currently available illusionist theories of consciousness cannot solve it.

3. The illusion meta-problem

A good way to explain the nature of the illusion meta-problem is to start with what many philosophers take to be the main flaw of illusionism: the fact that it is wildly implausible. To many, denying the existence of phenomenal states seems downright crazy. Chalmers thus writes that this kind of theory "denies the evidence of our own experience. This is the sort of thing that can only be done by a philosopher" (Chalmers, 1996, p. 188). Searle writes about Dennett's illusionism that it "denies the existence of the data"; and he comes close from calling this kind of position insane: "surely no sane person could deny the existence of feelings" (Searle, 1997). Other examples of such statements abound.[7]

I think that this radical implausibility of illusionism is one of the main reasons why this doctrine encountered only a limited success amongst people working on consciousness. Illusionism may have a lot of theoretical advantages, but it simply seems crazy. Many people think that asserting that consciousness is an illusion is a wildly puzzling, almost absurd claim. Illusionism easily triggers responses such as: "How could it be an illusion, as it is *right here*?"; "If I know something, I know that I am conscious right now"; "The idea that consciousness is an illusion is absurd, and no one seriously believe it. When you say that consciousness is an illusion, do you even *think* that it is true, or do you just *say* it?"

So, the situation is as follows: many people find that the idea that phenomenal consciousness is an illusion is very implausible and puzzling, and in fact borderline absurd. Some people consider that this is a sufficient reason to reject illusionism. Illusionists then try to meet these criticisms, either by attempting to show that illusionism is not as implausible as it may seem, or by insisting (rightly, in my view) that we should not rely too much on our intuitive sense of what is plausible to fix our philosophical beliefs.

My goal here is *not* to enter this first-order debate. I want to place myself on a second-order level: from an illusionist point of view, the fact that the illusionist hypothesis strikes many of us as wildly implausible, deeply puzzling, almost absurd, etc., is a fact about *the particular way in which we are subjected to the illusion of consciousness*. This shows that the illusion of consciousness is not like any old illusion. It is an illusion such that we face such a

[7] Bryan Frances writes, in a discussion on skepticism and error theories: "I assume that eliminativism about feelings really is crazy" (Frances, 2008, p. 241). Galen Strawson writes that eliminativists who deny that there are mental states such that there is something it is like to be in these states "do seem to be out of their minds", and that their position is "crazy, in a distinctively philosophical way" (Strawson, 1994, p. 101).

deep reluctance and puzzlement when we try to accept its illusory nature. I therefore think that the fact that most of us find illusionism wildly implausible, problematic, puzzling, etc., *is part of the explanandum* of illusionist theories. I will call the problem of explaining this fact the *illusion meta-problem*, and I will say that a theory solves the illusion meta-problem when it explains why illusionism strikes us as a wildly implausible, deeply puzzling, almost absurd thesis. I use the term "meta-problem", because this problem concerns the very acceptability of illusionism itself. However, I don't think that this problem is fundamentally *different* from the illusion problem. I rather see it as an aspect of the illusion problem – an aspect that has been seriously neglected until now.[8]

I think that solving the illusion meta-problem is essential for any illusionist theory of consciousness, as the radical implausibility of the illusionist thesis is part of the *explanandum* of illusionist theories. Moreover, I also think that, by solving this problem, illusionism could get some leverage in the first-order debate about its own plausibility that I mentioned earlier. Illusionists would indeed gain a simple answer to the reproach that illusionism is wildly implausible: "It is true that it is wildly implausible, but this implausibility is predicted by our theory. Therefore, it should not weigh against this theory if we have good independent reasons to accept it".[9]

One thing should be noted before we go further. I think that it is quite obvious that this implausibility of illusionism is deeply linked to the fact that we have a strong intuition that there is no appearance/reality distinction in the case of phenomenal consciousness. One can see that by considering the fact that one of the most immediate definition of "illusion" is: "a fallacious appearance", that is to say, an appearance of something that does not match the reality of the thing. If we think that there is no distinction between appearance and reality when it comes to phenomenal experience (and many people do, or at least find that position intuitively attractive), then we will think that there can be no illusion of phenomenal experience. However, I want to make it clear that, by insisting that illusionist theories have to solve the illusion meta-problem, I am of course *not* saying that they should explain that there *really is* no appearance/reality distinction when it comes to consciousness. So, one can recognize the need to solve the illusion meta-problem without endorsing the thesis that there

[8] By giving this aspect of the illusion problem a name of its own, I am aiming at attracting more attention to it.
[9] This is not to say that everyone would be satisfied by such an answer (Chalmers, 1996, p. 188‑189). However, I think that it is undeniable that it would give *some* dialectical force to the illusionist position.

is no appearance/reality distinction about consciousness.[10] However, I do think that illusionist theories who want to solve this problem will have to accept that we have an extremely strong intuition that there is no appearance/reality distinction about consciousness, and that this no appearance/reality distinction intuition is of such a kind that it makes illusionism about consciousness extremely implausible.[11]

Let's sum things up. While the illusion problem is the general problem of explaining why phenomenal consciousness appears to us to exist even though it does not, the illusion meta-problem is the problem of explaining why we are so deeply reluctant to accept that phenomenal consciousness is an illusion – why this idea strikes us a crazy, preposterous, and incoherent in a sense. The illusion meta-problem is a *part* of the illusion problem. Indeed, fully solving the illusion problem means explaining not only why we have the illusion that phenomenal consciousness exists, but also why we have *this particular kind of illusion*, with this particular strength. And it is part of the peculiar strength of the illusion of consciousness that we are deeply reluctant to accept that it is an illusion. In the next section, I will argue that current illusionist theories of consciousness (focusing on Pereboom's and Graziano's theory) cannot solve the illusion meta-problem.

4. Why current illusionist theories cannot solve the illusion meta-problem: the problem with Graziano's theory

I am now going to argue in a detailed manner for the thesis that current illusionist theories of consciousness cannot solve the illusion meta-problem. However, I won't make a systematic critical review of *all* current illusionist theories of consciousness – such a review would obviously be beyond the scope of this paper. I will rather focus on the two theories previously presented. I also want to make it clear that I don't intend to give a *principled* argument against *all* forms of illusionism on the subject of the illusion meta-problem. Indeed, I don't think that

[10] This point is important, as illusionists already seem to have some solid arguments against the thesis that there is no appearance/reality distinction when it comes to consciousness (Frankish, 2016, p. 32‑34). In fact, it should of course be noted that in the most obvious sense illusionist *must* deny that there is no appearance/reality distinction when it comes to consciousness, given that they say that conscious appears to exist without existing. I will return to this point in my concluding remarks.

[11] This should make clear that illusionists cannot solve this problem simply by interpreting the absence of appearance/reality distinction in a "weak" and "innocuous" way, so that they can endorse this absence of distinction themselves (Dennett, 1991, p. 81; Frankish, 2016, p. 33). Indeed, an absence of appearance/reality distinction which is innocuous enough to be compatible with illusionism cannot explain the resistance to illusionism. What illusionists have to explain is why we have a (fallacious) intuition of an absence of appearance/reality distinction in a "strong" (and not innocuous) sense of this distinction, so that we encounter this deep resistance to illusionism.

no illusionist theory can solve the illusion meta-problem – actually, I am quite convinced that some of them *can*. I simply think that currently available illusionist theories cannot solve this problem, unless they are seriously supplemented.

I will start by discussing what I take to be the most obvious way, for current illusionist theories of consciousness, to solve the illusion meta-problem – which is also the way used in Graziano's theory (for a first approach, Kammerer, 2016, p. 129). I will show why it fails to solve the problem. I will then turn to some more sophisticated considerations, that can be found in Pereboom's work, and which can be interpreted as ways to tackle the illusion meta-problem.

The most obvious thing illusionists could say concerning the illusion meta-problem amounts roughly to the following. In virtue of some hard-wired features of our introspective devices, we are disposed to believe that we are in states endowed with a peculiar qualitative nature – phenomenal states. And this disposition to believe is so strong, so pervasive, that we are very reluctant to accept any view that denies that we are in such states: we tend to judge that views of this kind are very implausible, even when we are provided with theoretical reasons to embrace such views. So, all in all, in this kind of account, the illusion meta-problem is solved simply by positing that the disposition to believe that we are in conscious states, caused by the peculiar nature of our introspective devices, is so strong that any view that contradicts what this disposition disposes to believe is considered to be wildly implausible.

I think that it is quite clear that Graziano endorses an account of that type. When it comes to explaining why his own position concerning consciousness seems implausible, Graziano draws a comparison with Newton's theory of colors (Graziano, 2013, p. 49, 80). Before Newton, people thought that white light was "pure", simple, not composed of any other colored lights. According to Graziano, this belief stems from the fact that our visual system encodes white as being a simple, primitive property. In virtue of this (hard-wired) feature of human perceptual devices, people had a strong disposition to believe that white light was pure and not composed of other colored lights. And when Newton first claimed that there was no such thing as a pure and simple white light, and that white light was in reality a mixture of all other colored lights, this view struck them as very implausible. Graziano thinks that the same thing is happening in the case of consciousness: our introspective mechanism (that Graziano calls the attention schema) represents our internal states as consisting in a simple relation of awareness, which is "a fluidic substance", an "experience", a "sentience".

And views that state that there are no such things in reality, but only sets of complex attentional processes, strike us as implausible.

I think that this kind of explanation fails to solve the illusion meta-problem, because it fails to account for the degree to which illusionism seems implausible to us. One could try to show it by pointing out the fact that illusionism concerning consciousness is so much more implausible that Newton's view of colors, even at Newton's time. However, I won't build my argument on that particular example, as in that case I would have to back it with some substantial historical knowledge. I will rather use another example, on the basis of which I will draw a systematic comparison in order to show why one cannot solve the illusion meta-problem simply by positing a strong introspective disposition to believe that we are phenomenally conscious.

Consider the following fact: it seems to me that I have two hands.[12] I have a strong disposition to believe that I have two hands, on the basis on visual perception, proprioception, memory, testimony, etc. This disposition certainly is very strong (probably even stronger that the disposition to believe that white light is pure), and it would take some serious evidence to convince me that *it is false that I have two hands right now*. Any view that implies that I do not have two hands right now will strike me as quite implausible, and I won't endorse it if it is not backed with robust evidence. However, the thesis according to which *I do not have two hands* does not strike me as puzzling or absurd. I have no difficulty entertaining the idea that *maybe* I do not have two hands; I have no problem picturing *what it would mean* for me not to have two hands. For example, I easily grant that I could be an armless person currently dreaming, or subject to a hallucination; or that I could be a brain-in-a-vat fooled by an evil scientist. All in all, I grant that it could very well *seem to me* that I have to hands while in fact I don't – though I don't believe this is the case, and I have a strong perceptive disposition to do so.

Now consider the fact that it seems to me that I am having a visual experience of the red dress of my friend who is sitting in front of me. This disposition to believe is based on introspection, and it is certainly quite strong. However, this strength goes further than in the previous case. Indeed, it is not only that I will find any views according to which I am not really having this experience right now quite implausible if not backed by robust evidence; it is not only that I will strongly tend *not* to believe such a view. I also have real trouble simply *picturing what it would mean not to have this visual experience right now, while it seems to*

[12] Here I draw on Moore's famous proof of the existence of an external world (Moore, 1939).

me that I have it. The idea that I am not conscious but that it seems to me that I am conscious seems deeply elusive to me. I find it deeply puzzling, and almost absurd: when I focus on the fact that it *seems to me* that I am having this experience, I cannot fully separate it from the fact that *I do have this experience*, so that there seems to be incoherent in the hypothesis.

The contrast with the previous example, I think, is quite obvious. The hypothesis according to which I don't have two hands, in spite of what perception, memory, etc., tell me, appears to me as a perfectly clear and coherent possibility (I could be dreaming, hallucinating, etc.), even though my (perceptive, testimonial, memorial) dispositions to believe that I have two hands are so strong that I would require an outstanding amount of converging evidence to lead me to abandon this belief. The same way (to quickly go back to Graziano's example), it seemed to us that white light is pure, simple, primitive, non-composed, because that is how perception presents it to us. And some robust evidence was needed before we started to accept that white light really is composed of all the other colored lights – dispersion of white light through a prism being a good start. However, even before they came to *believe* it, I think that people could easily picture what it would mean for white light to be composed: it meant that white light was *not at all what it seemed to be*; that it was not at all as we visually experienced it.

The situation is quite different in the case of consciousness. The hypothesis according to which I am not having this experience of my friend's red dress right now, even though it seems to me that I am having it, does not appear to me as a clear possibility. I am not sure I know what it would mean for this hypothesis to be true; I am not only reluctant to *believe* it to be true, I also find that there is something intrinsically troubling or even incoherent in this very hypothesis.[13] When I am thinking about this experience of a red dress as being only an illusion, I try to think of the *illusion* – of the fact that it seems to me that I have an experience of a red dress – and this introspective illusion seems enough to constitute the experience itself, which means that it is not really an illusion. This is the problem I encounter when I try to think that phenomenal consciousness is an illusion, and I encounter a similar problem in no other case of illusions. My opinion is that this problem comes from the deep difficulty we encounter when we try to draw a distinction between appearance and reality about phenomenal consciousness. However, the point I want to make here is more modest: I simply want to point out that illusionism regarding consciousness is much more implausible,

[13] Of course, I set aside here the question of knowing whether or not there is a sense in which this hypothesis really is incoherent (and why).

puzzling, and hard to represent to ourselves, than illusionism regarding any other thing (and even regarding things we are very strongly disposed to believe exist).

For this reason, Graziano's explanation of our resistance to illusionism is not satisfying, because it fails to do justice to the peculiar force of this resistance. However, one might try to strengthen Graziano's view by supplementing it. For example, one could say that our reluctance to envision the truth of illusionism can be explained if we grant that the thesis which asserts the reality of phenomenal states was indeed a thesis we came to believe because of what our introspective mechanism disposes us to believe, but that it then started to constitute a kind of contextually *a priori* statement within the study of the mind.[14] The idea of "contextually *a priori* statements" was introduced by Hilary Putnam (Ebbs, 2005; Putnam, 1975, 1983). This expression refers to statements which are *a priori*, but only in a way which is relative to a certain body of knowledge. Such statements are *a priori* relative to a body of knowledge, because within this body of knowledge we are not able to specify determinate ways in which these statements could be false. Contextually *a priori* statements are not rare or peculiar; history abouds with examples of them. The thesis according to which real, physical space is Euclidean was, during the 18th century, a typical example of a contextually *a priori* statement (the development of non-Euclidean geometries during the 19th century and the Einsteinian revolution, at the beginning of the 20th century later led us to recognize that it is in fact false). People at the time would have been deeply puzzled by someone telling them that external, physical space is not Euclidean – maybe even more than when they were told that white light is not pure, simple and non-composed. Indeed, before the development of non-Euclidean geometry, they simply weren't able to clearly picture the alternatives.

One could then try to use this concept and the corresponding examples to analyze the case of phenomenal consciousness. Maybe, for example, our introspective system represents our internal states as consisting in simple relations of awareness, which creates a strong disposition to believe that we are in such simple, primitive conscious states – this is what Graziano suggests. And maybe this strong belief is then "solidified", as it plays the role of a contextually *a priori* statement within the study of the human mind: we have no way to specify determinate ways in which this statement could be false, which makes us find its falsity inconceivable until theoretical breakthroughs provide us with plausible theoretical alternatives.

[14] I want to thank an anonymous reviewer for suggesting this explanation.

This explanation of our reluctance to accept illusionism is not exactly Graziano's, but it relevantly extends and strengthens Graziano's explanation with the help of Putnam's ideas. However, I don't think it provides a satisfying explanation of our peculiar resistance to illusionism regarding consciousness. Indeed, there is still a stark contrast between the difficulty people have when they consider the possible falsity of contextually *a priori* statements, and the puzzlement we encounter when we try to envision the truth of illusionism.

It is true that the falsity of a contextually *a priori* statement such as "real, physical space is Euclidean" seemed in a way inconceivable at a time, in the sense that people were unable to represent, in a determinate manner, ways that the world could be which would make this statement false. However, it is still obvious that people had *no difficulty* representing *in a partly indeterminate way* such a situation (and even believing it to be true). For example, it was quite "easy" for people of the 18th century to think that maybe the external reality *seemed to be structured in a spatially Euclidean way*, but that in fact this Euclidean spatial structure merely concerned the way in which reality *appeared* to us as thinking subjects, while reality in itself was maybe *not* so structured. After all, that is roughly what Kant (often considered as a typical example of a thinker who asserted that the Euclidean nature of physical space was *a priori*) argued in the *Critique of Pure Reason*, as he explained that the Euclidean spatial structure concerns the way in which things appeared to us (the phenomenon), but that we had no reason to believe that it also concerns the thing-in-itself. So, the falsity of the thesis according to which reality ultimately had an Euclidean spatial structure was quite "easy" to envision, even though people were not in a position to formulate precise and useful rival alternatives (such as hypotheses about the spatial, but *non-Euclidean* structure of reality), but only to represent this falsity in a rather indeterminate way.

By contrast, consider again the difficulty we encounter when we consider illusionism. In the case of consciousness, even a partially indeterminate understanding of illusionism appears deeply puzzling. When I consider the proposition "it seems to me that I am having this experience right now, but I am not really having it", what appears deeply problematic is the idea itself, quite apart from and prior to any consideration of the specific possible situation which would make this statement true. In the case of illusionism regarding consciousness, there seems to be something problematic and even incoherent in the hypothesis itself.

For this reason, I think that it is impossible to account for the peculiar resistance we encounter when we try to think that phenomenal consciousness is illusory simply by supposing that the reality of phenomenal consciousness plays the role of a contextually *a*

priori statement in our study of the mind. As I previously said, I think that deep down what is special in the case of consciousness stems from the peculiar difficulty we encounter when we try to draw a distinction between appearance and reality about it (which we don't encounter about any other entity). What I mostly want to point out now, though, is that the intuitive resistance encountered by illusionism is radically peculiar, and that this peculiarity is not explained by Graziano's theory (or by the modified version of it I just presented).

So, I think that one cannot explain our reluctance to accept illusionism regarding consciousness simply as a consequence of a strong introspective disposition to believe that we are conscious. And we cannot escape this difficulty by supposing that this strong disposition to believe has been solidified by the role played by the corresponding belief in our body of knowledge and our inquiries (appealing to Putnam's idea of contextually *a priori* statements). There is something specifically troubling in the very idea that consciousness is an illusion, which we don't encounter in the case of other entities.

I tried to present one way for the illusionist to explain why illusionism itself is so widely thought to be implausible. This explanation is the most natural answer an illusionist would give to the illusion meta-problem. This answer has been suggested by Graziano, and I think that many illusionists would naturally endorse it. I showed that it didn't account for the peculiar way in which illusionism is problematic for us. I argued that a refined version of this explanation (supported by Putnam's idea of contextually *a priori* statements) could not account for it either. The conclusion is that Graziano's theory (as well as refined version of this theory, or other theories which use the same kind of explanation[15]) cannot solve the illusion meta-problem.

5. Why current illusionist theories cannot solve the illusion meta-problem: the problem with Pereboom's theory

Graziano's strategy to account for the implausibility of illusionism is, I think, the most natural strategy for an illusionist. I tried to show why it fails. Pereboom's strategy is more complex and sophisticated. It relies on two distinct considerations. I will present these two considerations, and in each case I will argue that it cannot account for the peculiar difficulty we undergo when we try to think that phenomenal consciousness is illusory.

[15] I briefly argued elsewhere (Kammerer, 2016) that Humphrey's theory (Humphrey, 2011) falls in that category.

First, Pereboom may try to deflate the problem, by showing that the idea according to which introspection is inaccurate is not particularly counter-intuitive. Now, if the inaccuracy of introspection is not particularly counter-intuitive, then maybe its *systematic inaccuracy* is not particularly counter-intuitive either. And maybe the thesis that introspection systematically represents us as instantiating phenomenal properties that we don't really instantiate will not strike us as so implausible. Pereboom primarily makes this argument in the context of a first-order debate which directly bears on the plausibility of his own qualitative inaccuracy hypothesis. However, I think it could be used by a defender of his theory in the second-order debate which interests us here. After all, if illusionism is not as implausible, puzzling, etc., as I just stated, then maybe there is no actual illusion meta-problem.

Pereboom's strategy to show that we can intuitively admit the inaccuracy of introspection relies on examples. However, I think that these examples don't really show that illusionism is not radically implausible. First I will describe one of these examples, and then I will show why it does not help the illusionist who wants to deflate the illusion meta-problem.

Pereboom's first example (Pereboom, 2011, p. 22‑23), which is the one I will focus on, is borrowed from Christopher Hill (Hill, 1991, p. 128‑129), who borrowed it himself from Rogers Albritton. It involves a college student who is initiated into a fraternity. The student is shown a sharp razor, and is then blindfolded. At this point, he is told that the razor is about to cut his throat. He then feels something on his throat, and believes that it is *pain*. He then realizes that it is only a sensation of cold, and that an icicle had been put on his throat instead of a razor. Pereboom suggests that this example shows an intuitive example of *introspective inaccuracy*: for one second or so, the student *introspected pain*, rather than his real sensation – cold. One could use that example to argue that the intuition that introspection is infallible is not so strong after all, as we all admit that the case described is perfectly plausible.

However, I don't think that this example allows one to successfully deflate the *explanandum* for illusionist theories, and thus suppresses the illusion meta-problem. In a nutshell, I think that this example may convincingly show that we intuitively accept that *introspective judgments* can be false; however, it does show that we intuitively accept that *introspective appearances* can be fallacious. However, this is what Pereboom would need in order to deflate the illusion meta-problem.

In order to make my point, I will start by introducing the distinction between false judgments and fallacious appearances in the case of perception, and then turn this distinction to introspection. Let's say that I am in a room, in which there is a white table. It is possible

that the table *appears white* (perceptually), and yet that I mistakenly judge that it is, say, yellow. For example, I may have false beliefs concerning the lightning of the room. Or maybe I am expecting a yellow table in this room, because the table in the room has always been yellow, and I am not paying attention enough to the way in which the table actually appears to me, so that I mistakenly judge, in spite of the appearances, that the table is yellow. In that case, my *judgment* is mistaken, while the way in which the table appears to be is perfectly correct and accurate. This case is very different from a case in which the table *perceptually appears yellow* while it is white (because of the particular lightning condition, or because I am wearing special lenses, or because my visual system dysfunctions, etc.) – which would be a case of genuine fallacious appearance. And, if I judge that the white table is yellow *because it perceptually appears yellow to me*, then my judgment is defective because I have a perceptive fallacious appearance, and I commit a very different kind of mistake. So, it is one thing to form a mistaken judgment, and it is another thing to be presented with a fallacious appearance. This distinction echoes Christopher Hill's distinction between *errors of judgment* (wrong judgments made on the basis of correct appearances) and *errors of ignorance* – wrong judgments made on the basis of fallacious appearances (Hill, 1991, p. 127‑128; Pereboom, 2011, n. 33).

We can now go back to Pereboom's example with this distinction in mind. I think that his example can be convincingly interpreted as a case in which the student *experiences* cold, but *falsely judges* that he experiences pain. To that extent, the example may show that we are not strongly reluctant, from an intuitive point of view, to admit that introspective judgments may be false. This corresponds to Christopher Hill's interpretation of the example: according to Hill, in the fraternity case, the student forms a wrong judgment, but this is simply due to the influence of expectation upon his judgment – he commits an error of judgment, but no error of ignorance (actually, according to Hill, there can be no misleading appearances in the domain of introspection). However, Pereboom needs more than an intuitive description of a false introspective judgment in order to deflate the illusion meta-problem: he needs to show that it is plausible to interpret the case as a case in which the student forms his wrong judgment on the basis of a *fallacious introspective appearance*, so that he commits an *error of ignorance*. Indeed, illusionism is not simply the thesis according to which we *judge* phenomenal states to exist while they don't[16]; it's the thesis according to which phenomenal states *appear* to exist

[16] That would simply be "mistakism" (Prinz, 2016, p. 194), not illusionism – which is not to say that such a thesis would be devoid of interest in itself.

while they don't. Deflating the illusion meta-problem therefore requires showing that we are not deeply reluctant to accept that introspective *appearances* may be fallacious.[17]

Does Pereboom succeed in showing that this interpretation of the student example is plausible – or at least that we don't have a strong intuitive resistance against it? It is quite doubtful. His description of the example tends to avoid the contentious point by favoring neither the term "judgment" nor "appearance", but the more neutral term "representation"[18], even though at the same time he recognizes that he needs to interpret the example as an intuitive example of incorrect introspective *appearance*.[19] But given that *both* judgments *and* appearances could be seen as consisting (at least partly) in representations, the fact that Pereboom's example is an intuitive case of an incorrect introspective *representation* is not enough to show that it is an intuitive case of fallacious introspective appearance.

More importantly, I think that, when we try to examine Pereboom's example while keeping in mind the distinction between introspective judgments and introspective appearances, we are strongly intuitively pulled away from the interpretation Pereboom needs. In order to make this point clear, let us return to the fraternity example, and let us stipulate that the introspective *appearance* undergone by the student was indeed an introspective appearance of phenomenal pain. This means that, at least for a short amount of time, everything introspectively appeared to the student exactly as if he was experiencing pain. One consequence is that, had he been able to pay attention carefully to what introspectively appeared to him, his introspective judgment would not have varied, and he would still have judged, on the basis of what introspectively appeared to him (at least for this short moment), that he was experiencing pain. This specification is crucial, if we want to interpret his error as an error of ignorance, not an error of judgment.

[17] The concept of "introspective appearances" may itself appear problematic to some readers, given that many theories of introspection do not allow for such a thing – for example, theories which see introspection as a kind of *acquaintance*. For views of this kind (about which I will say a few words in conclusion), in introspection we are in a kind of direct relation to our phenomenal states, and there is no possible distinction between introspective appearances and introspected states themselves – in introspection, phenomenal states directly present themselves to the subject, without any intermediary. However, illusionists such as Pereboom have to refuse acquaintance views of introspection, as their thesis requires that phenomenal states only *indirectly appear* to the subject through introspection, so that they can introspectively *appear* to exist without really existing. Moreover, Pereboom explicitly uses the term "appearance" when talking about the outcome of introspection (Pereboom, 2011, n. 33).

[18] See for example: "the fraternity pledge at first misrepresents the qualitative features of the sensation of cold he actually has as qualitative features of pain" (Pereboom, 2011, p. 22).

[19] See what Pereboom says in a note: "such beliefs [false introspective beliefs] would be based on appearances that fail to do justice to the real qualitative nature of those properties [phenomenal properties]" (Pereboom, 2011, n. 33).

But, when we describe the case thus, I think that we are strongly led to judge that the student was really in fact, for a short moment, experiencing pain: he was really feeling the phenomenal quality of pain. If indeed everything appeared to him as if he was experiencing pain; if indeed, had he been able to carefully and attentively form his introspective judgment on the basis of his introspective appearance, he would have formed the judgment that he was in pain, then I think that we will strongly tend to say that there *really was* the phenomenal quality of pain in his experience. It is true that his experience of pain had neither the typical causes, nor the the typical effects associated with pain experiences, and that it lasted for a surprisingly short amount of time; but I think that we will be led to say that it had the phenomenal, qualitative *feeling* of pain nevertheless.

So, I think that Pereboom's example is inconclusive, because it does not manage to show that we are willing to accept that introspective appearances can be deceiving. It certainly shows that we intuitively accept that some of our introspective *judgments* can be wrong, but Pereboom needs considerably more than that. His description of the example relies on the use of a term which is neutral between judgment and appearance ("representation"); but nothing in his description makes a convincing case for the idea that his example is a plausible example of fallacious introspective *appearance* (as opposed to a mere case of wrong introspective judgment). Moreover, I think that, when we explicitly describe the fraternity case as a case in which the *introspective appearance* (and not only the introspective judgment) of the student is an introspective appearance of phenomenal pain, our intuition strongly leads us to the idea that the student *did feel phenomenal pain* in that situation, even though it was for a very short amount of time. But this speaks against the idea that we can intuitively recognize cases of discrepancies between phenomenal appearances and phenomenal reality. So, Pereboom's example does not conclusively show, in my opinion, that we can intuitively accept that there are fallacious introspective appearances. In conclusion, I think that Pereboom's first possible move – deflating the *explanandum* in order to suppress the illusion meta-problem – does not succeed.

However, Pereboom suggests a second move that could allow him, not to *deflate*, but to *solve* the illusion meta-problem. Indeed, he puts forth a hypothesis concerning the reason why "we are resistant to the possibility of such qualitative inaccuracy" (Pereboom, 2011, p. 23-24).[20] This hypothesis bears on a contrast between sensory representations and introspective

[20] "The qualitative inaccuracy hypothesis" is the name he gives to his own version of illusionism. The fact that Pereboom feels the need to *explain* the remaining resistance to the qualitative inaccuracy hypothesis suggests

representations. In the case of sensory representations, "we have readily available ways of checking the entity represented that are independent of the representation under scrutiny, while for introspection such ways of checking are at best very limited". For example, in the Müller-Lyer illusion, we perceptually represent two equal lines as being of *different lengths*, but we then can easily proceed to an objective measurement (using a ruler) and check that their lengths really are equal. However, nothing similar is possible (yet) in the case of introspection. This, according to Pereboom, explains why "we are at most only infrequently aware of discrepancies between the real qualitative natures of phenomenal properties and how they are introspectively represented", which in turns "provides a fairly plausible account of our resistance to the possibility of qualitative inaccuracy" (Pereboom, 2011, p. 23). So, to sum up Pereboom's explanation of our peculiar resistance to illusionism regarding consciousness: we have (at best) only rare opportunities to notice a discrepancy between introspective appearances and the reality of our phenomenal experiences, so we tend to judge that there are no such discrepancies. We are therefore led to think that introspective appearances are never fallacious, which in turn may explain why illusionism strikes us as wildly implausible and deeply puzzling.

I think that this hypothesis fails to explain the kind of intuitive resistance we undergo when we try to accept illusionism. I want to show it by way of a thought experiment.[21] Let's suppose that next year theoretical physicists come to suggest a new physical theory, the T theory, which proves very empirically successful. Because of this success, T ends up being adopted by the scientific community. Let's suppose that T posits the existence of a new kind of particles, epsilon particles. We currently cannot detect epsilon particles, but the theory tells us how to build a detector. Let's say that such a detector requires the use of a huge $20m^3$ cube of pure gold (more or less the total quantity of gold on Earth). Physicists manage to find enough money to fund the detector: they collect all the gold on Earth and they finally build the instrument. Of course, they can only build one. So, whenever the detector says "now there is a epsilon particle there", it is the case that physicists have *no independent way to check whether or not this is accurate*, as this detector is the *only instrument that can detect epsilon particles*.

The question I now want to ask is the following. Let's say that I am one of the physicists using this unique detector. Can I easily entertain the hypothesis that this detector is inaccurate

that he is not fully convinced himself that we are in fact ready to intuitively recognize the potential fallaciousness of introspective appearances (in spite of what his use of the fraternity example attempted to show).

[21] I presented a first version of this thought experiment, without the subsequent modifications and variations, in (Kammerer, 2016, p. 130).

– and maybe, even, *systematically inaccurate*? Would I find this hypothesis widely implausible, deeply puzzling, difficult to make sense of, or borderline incoherent? In my opinion, it is obvious that this would not be the case at all – not in the slightest. I think that I would have no trouble envisioning that this detector may be completely inaccurate – for example, maybe because the T theory is not accurate in spite of its empirical success, or maybe because someone made a mistake in the construction of the detector. The fact that this detector *is the only detector available when it comes do detecting epsilon particles* is completely irrelevant here. And if I come to have reasons to believe that this detector is inaccurate (for example, because I realize that a mistake has been made during the construction of the detector), then I will easily accept that it is likely that the detector systematically misrepresents the presence of epsilon particles, even if I have no independent way to check the presence or the absence of these particles.

This thought experiment presents a situation that is analogous to our own situation concerning consciousness according to Pereboom. Pereboom suggested that this situation may explain why we are so reluctant to embrace illusionism regarding consciousness. However, the conclusion of the thought experiment is that this explanation is unsatisfying. Indeed, if what caused our deep reluctance to accept illusionism concerning consciousness (the thesis that introspective appearances are systematically misleading) was the fact that we have no independent way (outside of introspection) to check whether our introspective representations are correct, then we should expect that physicists using the detector would encounter the same deep reluctance to accept that the hypothesis according to which the detector is systematically inaccurate. But this is obviously not the case: these imaginary physicists could easily accept that their detector is inaccurate, even without any independent way of checking for the presence of epsilon particles, provided they have some reasons to believe in the detector's inaccuracy (and these reasons may be purely theoretical reasons, and may have nothing to do with an independent observation of epsilon particles). And, even if they didn't *believe* their detector to be incorrect, they could surely easily *envision* the hypothesis that it is incorrect. So, Pereboom's hypothesis cannot explain the peculiar reluctance we encounter when we try to accept illusionism. The conclusion is that Pereboom's second move cannot solve the illusion meta-problem.

Joseph Levine, in conversation, made an objection to the reasoning I just presented. Maybe, he said, it is true that we would easily envision the idea that the detector I described on the thought experiment is inaccurate, even though we have no independent check on its

accuracy. However, it may be simply because we have lots of experience in general with physical object detectors of various kinds and understand the ways in which they can go wrong. But this is not the case with introspection, as we have no real idea how introspection works. This could explain why we are so reluctant to accept that introspective appearances may be misleading, and therefore why we encounter such a peculiar resistance to illusionism.

However, I think that this objection can be answered. First of all, it should be noted that Levine's objection does not allow for a defense of Pereboom's view strictly speaking. Indeed, Pereboom's view was that the cause of our reluctance to accept the idea that introspective appearances may be misleading is the fact that we have no independent check on the accuracy of introspection. I tried to show that this was probably false, as the fact that we have no independent check on the accuracy of a given source of information does not seem sufficient for us being deeply reluctant to envision that this source of information may be misleading. So, Levine's objection amounts to suggesting a slightly different, more sophisticated view, in which the cause of our reluctance to accept the idea that introspective appearances may be misleading is the fact that we have no independent check on the accuracy of introspection *together with* the fact that we have no clear idea of how introspection works *and* with the fact that we are not certain that introspection consists in the physical detection of a normal physical object. However, I don't think that such a view can account for the peculiar resistance we encounter when trying to accept that introspective appearances may be misleading. I will try to show that it fails, using another thought experiment.

Let's suppose that technologically and scientifically advanced aliens from Deneb come to visit Earth peacefully. Denebians spend time on Earth, and they tell us that they have a theory of the world (let's call it the U theory) which is much better than our physics, and which is so different from our physics that it could not properly be called a "physical" theory (as it is something quite distinct). However, they also refuse to explain to us the details of the U theory, and they simply tell us that this theory notably posits the existence of some spatio-temporally localized entities, amongst which are the "delta entities". They don't tell us exactly what these entities are, even though they may tell us about some very vague structural properties of these entities (such as the fact that they are spatio-temporally localized). When they leave Earth, we then realize that they forgot on Earth an object which we know is a detector of delta entities. We know how to turn on the detector, and we know when it "says" that there is a delta entity here and now. However, we have absolutely no idea how it works, and we have no clear idea of the real nature of delta entities.

Let's suppose we then start to use this detector to try to reconstruct the U theory (or an equivalent of the U theory). We will probably trust the detector very much, as it will be our *only reliable access* to delta entities. However, throughout our research, will we be able to *envision* the hypothesis that maybe this detector is inaccurate? I take it to be obvious that we will. And, for sure, the human physicists who will be trying to reconstruct the U theory, relying on this detector, will sometimes (in their moments of despair) have thoughts such as "Oh God, what if this detector is in fact inaccurate, or what if it dysfunctions! Then most of what we do is useless and misguided, and our research is in a dead end". So, this shows that they will easily envision the hypothesis that the detector of delta particles could be inaccurate. But this detector, by hypothesis, is such that (1) we have no independent check on its accuracy, (2) we have no idea how it works and (3) we are not even certain that the particles in question are "physical", as the Denebians told us that the U theory was so different from our physics. So, I think that this thought experiment shows that the fact that a given source of information is such that (1) we have no independent check on its accuracy, (2) we have no idea how it works and (3) we don't properly think that what it detects is physical, is not sufficient to create a deep reluctance to clearly and simply *envision* the hypothesis that this source of information may be inaccurate. Therefore, I think that Levine's suggestion cannot account for the peculiar resistance we encounter when we try to envision the idea that introspective appearances may be misleading, and that illusionism may be true.

6. Concluding Remarks

In this paper, I described a problem that illusionist theories of consciousness have to solve, which I called the "illusion meta-problem". It is the problem of explaining why, even though illusionism is true, it strikes us as wildly implausible, deeply puzzling, almost absurd; why we are having such a hard time simply picturing what it means for phenomenal states to be illusory. I think that currently available illusionist theories of consciousness cannot solve this problem. I focused on two of the most promising illusionist theories (Pereboom's and Graziano's). I tried to show that these theories did not explain the peculiar difficulty we face when trying to accept illusionism. In Graziano's theory, the illusion of consciousness has the same status as any other illusion (such as perceptual illusions). In this view, the fact that the illusory nature of consciousness is so difficult to accept (much more than the illusory nature of other things) is left unexplained. Pereboom puts forth two considerations which are

relevant to this issue. One amounts to *denying* that there is something peculiar in the reluctance we face when we consider the illusory nature of consciousness. I tried to show that this move was unconvincing. The other one is a hypothesis that aims at explaining why there is after all a peculiar difficulty attached to the acceptance of the illusory nature of phenomenal states. I argued that this hypothesis was unable to explain the peculiarity of our intuitive resistance to illusionism.

My goal, in this paper, was not to argue against illusionism in general. I am not claiming that *no* illusionist theory of consciousness can solve the illusion meta-problem. On the contrary, I am quite convinced that there are such theories. In order for illusionists to solve this problem, I think that they must give more importance to the fact that we encounter such insurmountable difficulties when we try to distinguish between phenomenal appearance and phenomenal reality.

To a certain extent, the problem encountered by illusionists can be analyzed through the opposition between the two main available views of phenomenal introspection. Indeed, we can say (with a lot of simplification) that there are two main kinds of theories of phenomenal introspection in philosophy of mind. The first kind sees our own relation to our phenomenal states as something akin to a kind of *acquaintance* (for the first use of the term, see Russell, 1912), so that the introspected state is in a way a *part* of the introspective state. In this view, introspecting subjects are in the most direct relation to their phenomenal states. Views of this kind are often non-physicalist (for some recent versions of these views, see Gertler, 2012; Goff, 2015), but there have been recent attempts to give physicalist accounts of such a relation of *acquaintance*, or more generally of the idea according to which, in introspection, phenomenal states present themselves (Balog, 2012; Kriegel, 2009; Loar, 1997; Prinz, 2016). This kind of acquaintance view quite naturally explains why we face such difficulties when we try to distinguish between phenomenal appearance and phenomenal reality (and it therefore explains why illusionism seems so unbelievable), but only by actually *endorsing* the thesis according to which there really is no such difference. For this reason, such views are incompatible with illusionism.

The other kind of theories of introspection sees phenomenal introspection as something more akin to perception, in the sense that it considers that, in introspection, there is an actual distinction between the introspected state and the introspective state. In this view, introspective representations are not constituted (or partly constituted) by phenomenal states, which means the two can come apart. Monitoring theories of introspection (or "causal"

theories of introspection) typically fall in that category (Armstrong, 1980; Nichols & Stich, 2003; Pereboom, 2011).[22] This second kind of theories is compatible with illusionism. The problem is that this kind of view provides no natural explanation for the trouble in distinguishing between phenomenal appearance and phenomenal reality – which seems precisely to be what we have to explain in order to solve the illusion meta-problem. Views of this kind have to be supplemented in order to be able to provide such an explanation – but, as I tried to show in this paper, such a task is neither simple nor easy. Therefore, one way to describe the difficult theoretical situation which illusionists find themselves in is to say that illusionists face the following constraints: they have both to endorse a theory of introspection that sees introspection as something *fallible*, involving two distinct states (the introspective state and the introspected state) – that is to say, they have to deny that we are acquainted with our phenomenal states – in order to maintain that introspection is illusory, and at the same time they have to show why acquaintance itself seems so strongly and inescapably plausible as a theory of introspection.[23]

How might illusionists solve this difficulty? One hypothesis that, in my opinion, deserves to be explored is that introspection does *not* provide acquaintance with phenomenal states, and that there is a clear distinction between our introspected states and our introspective states; but that our introspection representations represent phenomenal states *precisely as having a peculiar epistemological nature*, and such that their appearance and their reality conflate. In this kind of view, introspection does not involve acquaintance with our phenomenal states, but it is loaded in such a way that our introspective representations falsely characterize our phenomenal states as states with which we are acquainted. This could be explained if we grant, for example, that introspection is partially determined by our naïve theory of mind, which notably features the naïve concept of "appearance" in its conceptual repertoire. I think that a view in the vicinity (and perhaps also other views) may be in a position to solve the illusion meta-problem. Some proposals in this direction can be found in (Kammerer, 2016). Whether or not a hypothesis of this kind can successfully solve the

[22] According to Pereboom, the first kind of theories can be tracked back to Brentano, while the second one is Kantian in inspiration (Pereboom, 2011, p. 3; 19).

[23] The following point is worth noting: even someone who thinks that illusionism is false still has to recognize that we face a very peculiar (and surprisingly strong) resistance when we try to envision its truth. Acquaintance theorists may have a natural way to deny illusionism *and at the same time* to explain this peculiar intuitive resistance, but this is not the case of phenomenal realists who endorse a causal model (non-acquaintance based) of introspection. Therefore, explaining this peculiar resistance is not only a difficulty for illusionists; it can also be seen as a difficulty for phenomenal *realists* who refuses to see introspection as a form of acquaintance (Armstrong, 1980; Nichols & Stich, 2003). The extent to which these theories of introspection, developed outside the context of illusionism, have the resources to solve this difficulty deserves to be explored – although it is beyond the scope of this paper. I want to thank an anonymous reviewer for this remark.

illusion meta-problem, my conclusion is that this problem has to be taken into account by illusionists, if they want to provide a fully satisfying theory of the peculiar way in which we are subjected to the illusion of phenomenality.

References

Armstrong, D. (1980). *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press.

Balog, K. (2012). Acquaintance and the Mind-Body problem. In C. Hill & S. Gozzano (Ed.), *New Perspectives on Type Identity: The Mental and the Physical*. Cambridge University Press.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200‑19.

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

Dennett, D. (1988). Quining Qualia. In A. Marcel & E. Bisiach (Ed.), *Consciousness in Modern Science*. Oxford University Press.

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

Ebbs, G. (2005). Putnam and the Contextually A Priori. In A. Randall E., A. Douglas R., & Lewis Edwin Hahn (Ed.), *The Philosophy of Hilary Putnam* (p. 389‑411). La Salle, Illinois: Open Court.

Frances, B. (2008). Live Skeptical Hypotheses. In J. Greco (Ed.), *The Oxford Handbook of Skepticism* (p. 225‑244). Oxford: Oxford University Press.

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

Gertler, B. (2012). Renewed Acquaintance. In D. Smithies & D. Stoljar (Ed.), *Introspection and Consciousness* (p. 89–123). Oxford University Press.

Goff, P. (2015). Real Acquaintance and Physicalism. In P. Coates & S. Coleman (Ed.), *Phenomenal Qualities: Sense, Perception and Consciousness*. Oxford University Press.

Graziano, M. (2013). *Consciousness and the Social Brain*. Oxford: Oxford University Press.

Graziano, M. (2016, january 12). Consciousness Is Not Mysterious. *The Atlantic*. http://www.theatlantic.com/science/archive/2016/01/consciousness-color-brain/423522/

Hill, C. (1991). *Sensations: A Defense of Type Materialism*. Cambridge: Cambridge University Press.

Humphrey, N. (2011). *Soul Dust: The Magic of Consciousness*. Princeton: Princeton University Press.

Kammerer, F. (2016). The hardest aspect of the illusion problem - and how to solve it. *Journal of Consciousness Studies*, *23*(11‑12), 123‑139.

Kriegel, U. (2009). *Subjective Consciousness : A Self-Representational Theory*. Oxford University Press.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, *64*(October), 354‑61.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford University Press.

Loar, B. (1997). Phenomenal States (Revised Version). In N. Block, O. Flanagan, & G. Güzeldere (Ed.), *The Nature of Consciousness* (p. 597‑616). MIT Press.

Moore, G. E. (1939). Proof of an External World. *Proceedings of the British Academy*, *25*, 273‑300.

Nichols, S., & Stich, S. (2003). How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. In Q. Smith & A. Jokic (Ed.), *Consciousness: New Philosophical Perspectives*. Oxford: Oxford University Press.

Pereboom, D. (2009). Consciousness and Introspective Inaccuracy. In L. Jorgensen & S. Newlands (Ed.), *Appearance, Reality, and the Good: Themes from the Philosophy of Robert M. Adams* (p. 156‑187). Oxford University Press.

Pereboom, D. (2011). *Consciousness and the Prospects of Physicalism*. Oxford University Press.

Prinz, J. (2016). Against Illusionism. *Journal of Consciousness Studies*, *23*(11‑12), 186‑196.

Putnam, H. (1975). It ain't necessarily so. In *Philosophical Papers, Volume 1* (p. 237‑249). Cambridge: Cambridge University Press.

Putnam, H. (1983). There is at least one a priori truth. In *Philosophical Papers, Volume 3*. Cambridge: Cambridge University Press.

Rey, G. (1995). Towards a Projectivist Account of Conscious Experience. In T. Metzinger (Ed.), *Conscious Experience*. Paderborn: Ferdinand Schoningh.

Russell, B. (1912). *The Problems of Philosophy*. Oxford University Press.

Searle, J. (1997). *The Mystery of Consciousness*. New York: The New York Review of Books.

Strawson, G. (1994). *Mental Reality*. Cambridge (Mass): MIT Press.