This is a last draft of an article accepted for publication and forthcoming in *Erkenntnis*. Please cite the final published version.

How rich is the illusion of consciousness?

Abstract:

Illusionists claim that phenomenal consciousness does not exist, but merely seems to exist. Most debates concerning illusionism focus on whether or not it is true – whether phenomenal consciousness really is an illusion. Here I want to tackle a different question: assuming illusionism is true, what kind of illusion is the illusion of phenomenality? Is it a "rich" illusion – the cognitively impenetrable activation of an *incorrect* representation – or a "sparse" illusion – the cognitively impenetrable activation of an *incomplete* representation, which leads to drawing incorrect judgments? I present this distinction and I classify the most influential illusionist theories along this line of divide. I then offer an argument against the accounts of the illusion of phenomenality in terms of sparse illusion.

Introduction[1]

Illusionism is the thesis according to which phenomenal consciousness is an illusion: it does not really exist, even though it seems to exist. Illusionism is very much a minority view: it is widely judged to be counterintuitive, and many philosophers find it implausible. However, this view has found some prominent defenders in contemporary philosophy.

Most debates regarding illusionism concern its plausibility as a theory of consciousness: is phenomenal consciousness really just an illusion? Here my goal is to address another

[1] I want to thank Keith Frankish and Sonia Paz Higgins for their comments and their help, as well as the audience at the University of Edinburgh (Designed Minds conference 2017), and three anonymous reviewers from their useful remarks.

question: *if* illusionism is true, that is, *if* phenomenal consciousness really is just an illusion, *what kind of illusion is it*? Such interrogation can itself lead to various more specific questions. I want to focus on one in particular: if phenomenality is an illusion, is it a *rich illusion* or a *sparse illusion*? By "rich illusion", I mean an illusion in which an object is *positively presented in an incorrect way* by a cognitively impenetrable representational process; by "sparse illusion" I mean an illusion in which the object is *not positively presented in an incorrect way* by a cognitively impenetrable representational process – but presented in a merely *partial* or *incomplete* way, which in turn leads us to *infer* incorrect beliefs about the object.

First, I will present illusionism about consciousness (§1), and the distinction between rich and sparse illusions (§2). I will then use this distinction to classify major illusionist views of consciousness: I will present what I call "rich-illusion views" and "sparse-illusion views" (§3). I will stress some of the advantages of the sparse-illusion views (§4). I will finally develop an argument against sparse-illusion views, which will also provide support for rich-illusion views (§5).

1. Illusionism about phenomenal consciousness

Phenomenal states (or "conscious experiences") are mental states which instantiate phenomenal properties (or "*qualia*"); in virtue of such instantiation, there is "something it is like" to be in those states. A visual sensation of green, an olfactory sensation of honeysuckle, the sensation of pain that we feel when the dentist touches a nerve, are supposed to be typical examples of phenomenal states. A creature is phenomenally conscious, i.e. possesses phenomenal consciousness, if and only if it enters phenomenal states.

Phenomenal consciousness seems to be a particularly mysterious aspect of our mental life. Indeed, most philosophers admit that phenomenal consciousness creates a *prima facie* difficulty

for the physicalist view of the mind, which is the dominant conception in the metaphysics of mind. Indeed, phenomenal states appear quite peculiar: they seem to consist in an irreducibly subjective apprehension of intrinsic and ineffable qualities. The idea that such peculiar states could be, in some sense, nothing over and above purely physical processes – as the physicalist says – strikes us as puzzling. It has been said that phenomenal consciousness creates a "hard problem" for physicalist views (Chalmers, 1995), or that there is a remaining "explanatory gap" between consciousness and the physical (Levine, 1983).

Illusionism (Frankish, 2016) states that phenomenal consciousness does not exist, even though it seems to exist: we never enter phenomenal states, even though it seems to us that we do. For the illusionist, none of our mental states ever instantiate phenomenal properties, so that, strictly speaking, there is *nothing it is like* to be in any of our mental states. Illusionists can still accept that the mental states we usually call "phenomenal states" have "quasi-phenomenal properties" (Frankish, 2016, p. 15): purely physical/functional properties of brain states which are reliably tracked but mischaracterized as phenomenal by our introspective representations.[2] But, crucially, quasi-phenomenal properties are *non-phenomenal* properties (contrary to what introspection presents), and there is nothing it is like to be in quasi-phenomenal states.

It is to be noted that Keith Frankish distinguishes between strong illusionism and weak illusionism (Frankish, 2016, p. 15-16). While strong illusionists claim that phenomenal consciousness does not exist, weak illusionists claim that phenomenal consciousness exists, but does not have many of the properties it seems to have. Here I focus on strong illusionism, that I simply call "illusionism" (following Frankish).

One of the main advantages of illusionism is that it avoids the hard problem of consciousness and dissolves the explanatory gap. We cannot comprehend how phenomenal

[2] Illusionism is compatible with any view regarding whether or not the set of quasi-phenomenal states forms a natural kind. It is also compatible with any view regarding the existence and the nature of other, non-phenomenal forms of consciousness – such as access-consciousness.

consciousness could be nothing over and above purely physical processes, and we cannot bridge the explanatory gap between phenomenal consciousness and the physical. However, if we embrace illusionism, this does not create any serious worry for physicalism, as phenomenal consciousness simply does not exist. As Keith 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). The illusion problem seems much more tractable, from a physicalist perspective, than the original hard problem. For this reason, illusionism provides a robust defense of physicalism regarding the human mind against the threat created by phenomenal consciousness.

In spite of this advantage, illusionism regarding consciousness certainly remains a minority view. It should be noted, however, that it is by no means a *marginal* view, as versions of illusionism (or neighboring views) have been developed and defended by many philosophers and scientists (Clark, 2000; Dennett, 1988, 1991, 2017; Drescher, 2006; Frankish, 2016; Graziano, 2013; Humphrey, 2011; Kammerer, 2016, 2019b, 2019c; Pereboom, 2009, 2011; Rey, 1995) amongst whom some are prominent thinkers in the field of consciousness studies. Daniel Dennett even called illusionism the "obvious default theory of consciousness" (Dennett, 2016). Illusionism also received praise from influential critics of physicalism ("if I were a materialist, I would be an illusionist" (Chalmers, 2018)).

Much of the philosophical discussion regarding illusionism so far has concerned its plausibility as a theory of consciousness: is consciousness really just an illusion? Is such a view likely to be true? Is it even entirely *coherent*? Isn't it *ad hoc*? (One can for example wonder: why on earth we should be the victims of such an illusion of phenomenality? Would this illusion have any kind of adaptive value?)

Here I want to focus on a different question: if phenomenal consciousness is an illusion, what kind of illusion is it? This question can of course receive many interpretations and give

rise to different debates, given that there are various ways to classify illusions. The question I want to focus on is the following: is the illusion of phenomenality a *rich* or a *sparse* illusion? I will now explain what I mean exactly by "rich" and "sparse" illusions.

2. Rich illusions and sparse illusions

First, what is an *illusion*? Here I will not distinguish it from a *hallucination*, and I will simply begin by saying that an illusion is an incorrect representation of a current situation: when one undergoes an illusion, one represents the presence of a feature X in a current situation S – when S does not really have the feature X – or the absence of a feature Y – when S does have the feature Y. However, defined in that way, there would be no difference between an illusion and a *false belief* about a current situation. One therefore has to add a key element: illusions have a certain degree of *cognitive impenetrability*, which can be understood as informational encapsulation relative to information stored in central memory – typically, under the form of *beliefs* (Pylyshyn, 1984). That is, illusions, at least to a certain extent, are resistant to the acquisition of new relevant information encoded in a belief-like format about the situation which is the object of the illusion. For example, take a standard example such as the Müller-Lyer illusion. When we are facing the two lines of the Müller-Lyer illusion, we visually represent the two lines as having different lengths, even when we *know* that they really have the same length. On the other hand, if we simply *believe* that we are facing two lines of a different length (without looking, for example), coming to know that the two lines really have the same length destroys that belief. So, illusions are different from false beliefs because of their superior degree of cognitive impenetrability.

However, depending on the precise degree of cognitive impenetrability (and on the *level* at which such impenetrability holds), we can distinguish between what I call *rich illusions* and

sparse illusions. When we undergo a *rich illusion*, we activate a representation of a situation S characterized as having X (if S does not have X) or as lacking Y (if S does have Y), and the activation of this representation happens through a fully *cognitively impenetrable process*: no acquisition of new information will in itself hinder this activation.[3] I call this kind of illusion "rich" because the presence of X (or the absence of Y) is included in the very content of the representation produced by the cognitively impenetrable process. The Müller-Lyer illusion is a rich illusion. The illusion created by looking at the hologram of, say, a cat, is also a rich illusion. In the first case, I activate a perceptual representation of two lines as having different lengths; in the second case, I activate a perceptual representation of a tri-dimensional cat being in front of me; and in both cases, this activation is cognitively impenetrable.

On the other hand, there are *sparse* illusions. In the case of a sparse illusion, we activate (in a cognitively impenetrable way) a representation of a situation S which is a sparse, partial and *incomplete* representation of S, but which does not represent the *presence* of an absent feature, or the *absence* of a present feature. This representation is not *incorrect*; however, this representation is partial and incomplete in such a way that we are led to *infer* (in a systematic, intuitive but *cognitively penetrable way*) the presence of some non-existent feature, or the absence of some existent feature. That is, this representation leads us very naturally to an *incorrect judgment*. I think that most illusions created by magicians, for example, are *sparse illusions*. For example, consider the "headless woman" illusion, described by David Armstrong:

[3] Here, I consider that "fully cognitively impenetrable" means "synchronically impenetrable" (acquiring new information at a single moment, as such, does not substantively impact the relevant representational process) *and* "diachronically impenetrable in adults" (acquiring new information at an adult age, as such, does not substantively impact the relevant representational process, even over an extended period of time). I set aside the problem of diachronic impenetrability through earlier stages of development in the context of this paper, and I do not require a representational process to possess this kind of diachronic impenetrability to qualify as "fully cognitively impenetrable". For the distinction between diachronic and synchronic penetrability/impenetrability, see (McCauley & Henrich, 2006).

> "To produce this illusion, a woman is placed on a suitably illuminated stage with a dark background and a black cloth is placed over her head. It looks to the spectator as if she has no head. The spectators cannot see the woman's head. But they gain the impression that they can see that the woman has not got a head. (*Cf.* 'I looked inside and I saw that he was not there'). Unsophisticated spectators might conclude that the woman did not in fact have a head.
>
> What the example shows is that, in certain cases, it is very natural for human beings to pass from something that is true: 'I do not perceive that X is Y', to something that may be false: "I perceive that X is not Y". (Armstrong, 1968, p. 48)

This is a typical example of a sparse illusion. Subjects facing a headless woman activate a cognitively impenetrable representation of the woman, which does not present her head. This representation is not incorrect, but it is *incomplete*. However, it is incomplete in such a way that it naturally leads the subject to infer that the woman really has no head – the subject is led to an *incorrect* judgment about the situation. But this inferential process is cognitively penetrable; if the subject is careful enough (for example because she has seen this illusion many times before and knows how it works), she can avoid drawing such a conclusion.

To use a metaphor, rich illusions are lies of commission: our perceptual system delivers an *incorrect* representation of the world. Sparse illusions are lies of omission: our perceptual system only delivers an *incomplete* (but correct) representation, and we make a (cognitively penetrable) *mistake* when we infer something incorrect about the world, on the basis of this incomplete representation. Coming to know that a rich illusion *is* a rich illusion, we come to know that we should not trust the verdict of our perceptual capacities. Coming to know that a sparse illusion *is* a sparse illusion, we come to know that, even if the verdict of our perceptual

capacities can be trusted, we should not trust the judgments that we instinctively tend to draw on the basis of our perception.

It should be clear that the distinction between *rich* and *sparse* illusions does not exhaust the logical possibilities here. After all, an illusion which does not consist in the cognitively impenetrable activation of an incorrect representation is not *rich*, but amongst *non-rich* illusions, only some of them are *sparse* – those which consist in the cognitively impenetrable activation of an incomplete representation, which tends (in virtue of its incompleteness) to trigger inferences to false beliefs. However, there are potentially illusions which are not *rich* (the *incorrect* representation is not delivered through a fully cognitively impenetrable process), but are not delivered either through a process in which an incomplete representation is first activated (in a cognitively impenetrable way) and an incorrect judgment is then drawn on the basis of this incomplete representation and in virtue of its incompleteness. For example, it could be because the incorrect judgment which tends to be drawn has not directly to do with the prior cognitively impenetrable representation being incomplete in specific ways, but rather with tendencies to judge incorrectly which are themselves caused by prior beliefs, values, etc. In other words, there is room in the logical space for illusions which are *neither rich nor sparse* – I will later mention potential views which imply that the illusion of consciousness falls in this category. However, I think that most (if not all) major illusionist theories of consciousness, as I will show, think of the illusion of consciousness either as a *rich* or a *sparse* illusion. I therefore suggest that we now turn to the question: is the illusion of phenomenality *rich* or *sparse*? [4]

[4] Of course, one could also want to reserve the term "illusion" for what I call "rich illusions" – for example because they think that the processes generating illusions have to be cognitively impenetrable all the way down (as a matter of definition). I do not want to enter in this kind of semantic debate here; most of what I am going to say could be understood and accepted by someone who has this kind of strict understanding of illusions, provided they engage in some renaming – instead of asking "is the illusion of consciousness rich or sparse?" they could ask "is consciousness an introspective illusion, or a mistake we instinctively tend to make on the basis of incomplete introspective representations?"

3. Rich-illusion accounts and sparse-illusion accounts

What would it mean, for the illusion of phenomenality to be a rich illusion? It would mean the following: when we introspect, we token cognitively impenetrable representations which represent the instantiation of non-existent features, or the non-instantiation of existent features. For example, it could be that we then represent our internal states as instantiating *phenomenal properties,* characterized positively as having features which are not instantiated in reality – as being inherently qualitative, immediately known, etc. – or as lacking features that all properties of our internal states really have – say, as lacking a physical nature. Such misrepresentation would arise at the very level of introspection, in a cognitively impenetrable way.

If the illusion of phenomenality is sparse, on the other hand, that means that, when we introspect, we activate cognitively impenetrable representations which are not *positively incorrect*. These representations merely represent our internal states in a sparse, incomplete and partial way. But this incomplete (or "schematic") representation is such that it leads us to *infer* (in a systematic, intuitive, but cognitively penetrable way) that our internal states instantiate non-existent features, or do not have some of their existent features – for example, it leads us to infer that some of our internal states really are *nothing more* than what our representations represent, and do not have any other properties (do not have physical properties, for example).

Where do the main current illusionist views of consciousness fall along this distinction? I take it that some illusionist theories are clearly *sparse-illusion* views. For example, Michael Graziano's "attention schema theory" seems to me to be quite a typical example of a sparse-illusion view.[5] According to Graziano (Graziano, 2013, 2016), our brain forms a schematic representation of its own attentional processes: the "attention schema". It is a simplified,

[5] Graziano rejects the term "illusionism" (Graziano, 2016) to qualify his theory, because he wants to limit the use of the vocabulary of "illusion" to rare and abnormal dysfunctions of a detecting mechanism. However, I think he clearly is an illusionist in the sense I have given the term.

incomplete and schematic representation (and is such as a result of a trade-off between accuracy and processing resources). Instead of representing attentional processes in all their complexities, it represents a simple relation of "awareness" between a subject and a piece of information. However, this representation is not positively incorrect: the problem is only that we tend to commit a mistake when we then judge that we really enter internal states which are as described by our attention schema *and do not have other properties* (for example, do not have a complex internal physical nature). Graziano draws a comparison between the case of consciousness and our intuitive conception of white light: our perceptual system presents the color white in a schematic and incomplete way, which does not present the fact that white light is composed by all the colored lights of the spectrum. For that reason, we are intuitively led to think of white light as actually *pure* and *primitive*; as *not composed by anything* – which is why, according to Graziano, most people intuitively rejected Newton's theory of colors when it was first suggested, as this theory states correctly that white light is a mixture of all other colored lights (Graziano, 2013, p. 49, 80). Graziano's account, therefore, clearly is a sparse-illusion view: introspection delivers *incomplete* representation, which then misleads us into forming *incorrect* beliefs. [6]

On the other hand, Pereboom's "qualitative inaccuracy hypothesis" (Pereboom, 2009, 2011) is a typical case of a *rich-illusion* view.[7] 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 other words, the qualitative nature of the phenomenal properties of our internal states is directly represented

[6] One other way to interpret Graziano's theory would be to say that our attention schema explicitly represents our internal states as *not having* any properties outside of the ones it ascribes them. That would make Graziano's view a rich-illusion account. However, I think that Graziano chooses the first interpretation, as he insists that the attention schema is merely *incomplete* and *schematic* (Graziano, 2016, p. 104) but not positively incorrect. Moreover, he has confirmed that this was his favored interpretation in conversation.

[7] Pereboom does not explicitly endorse this hypothesis in the book cited, even though he tries to make the case that it constitutes an open possibility. However, for reasons of simplicity, I will speak of this view as if Pereboom endorsed it.

by introspection (that Pereboom conceives of as a perception-like mechanism). The reason why introspection delivers such a verdict about the qualitative nature of phenomenal properties, according to Pereboom, is to be found in some hardwired features of our introspective mechanisms (about which he does not speculate). The crucial point here is that the process that leads to the activation of an incorrect representation of our internal states is not cognitively penetrable. No amount of reflection, no acquisition of new information, can change the fact that introspection will present the phenomenal properties of our internal states as having a qualitative nature that they lack in reality.

Dennett's view on consciousness is perhaps harder to classify, although I think it is more natural to interpret it as a sparse-illusion view. One of the earliest defenders of illusionism, Dennett has developed this perspective in many publications (Dennett, 1988, 1991, 2016). In his latest book he claims that consciousness is a kind of "user-illusion" (Dennett, 2017, Chapter 14; Frankish, 2016, p. 16-17), similar to the user illusions we may undergo when we use the graphical interface of a computer. The icons displayed on a computer screen are only a schematic, abstract and simplified representation of complex computational structures hosted in the computer: they allow for a simple manipulation of the computer even though they do not provide any understanding of its internal working. In the same way, when we form introspective representations of phenomenal states, what we get is nothing but schematic and simplified representations of complex brain states and structures. In both cases, an illusion arises as we tend to think that the simple entities we represent (the icons, or the phenomenal states) *are really there in reality, exactly as our representations depict them,* in the deep and ultimate structure of the system (the computer, or the mind). We are victims of a user-illusion when we think that there really are simple, basic and unstructured files (as unstructured as the schematic representation of them provided by the icons on our desktop). In the same way, we are victims of a similar illusion when we think that our minds are populated with simple, primitive and

basic phenomenal states (as simple and unstructured as what our schematic representations present).

The most natural interpretation of this view is to see it as a sparse-illusion view: introspection delivers (in a cognitively impenetrable way) a schematic and incomplete representation of our internal states, and we then make a mistake when we tend to infer (in a cognitively penetrable way) that our internal states really are as simple and unstructured as our representations of them. From the *absence of representation* of a complex internal structure, we conclude to the *absence* of complex internal structure. The fact that this last step consists of a cognitively penetrable inference (although it may be made in a systematic, habitual and non-reflective way in most cases) is crucial to interpret Dennett's view as a sparse-illusion view. I think this interpretation is made natural by the fact that Dennett talks about such an inference as resulting from a process through which we "involuntarily misinterpret" (Dennett, 2017, p. 176) our partial and incomplete introspective representations, as well as by the fact that he draws an explicit parallel with other mistakes that we make on the basis of sparse representations, and which are very likely to be cognitively penetrable (when we tend to posit that there have to be intrinsic properties such as "cuteness" of babies, "sexiness" of handsome persons or "funniness" of jokes, that account for our desires to cuddle, our lust or our laughing, instead of seeing these properties as complex dispositions (Dennett, 2017, p. 171-172)). However, interpreting this mistaken inference as a non-cognitively penetrable inferential process (which I think is not the natural interpretation here) would make Dennett's view a rich-illusion view.

Other illusionist views of consciousness can be classified along this line of divide – although I will not argue for my categorizations. Some illusionist views clearly are, I think, sparse-illusion views (Drescher, 2006; Rey, 1995; Shabasson, ms); others are clearly rich-illusion views (Kammerer, 2016, 2019b, 2019c); some, I think, can be interpreted in both ways

(Humphrey, 2011) – although I suspect that in Humphrey's case the most natural interpretation makes it a rich-illusion view.[8] [9]

4. Advantages of sparse-illusion views

Although currently available illusionist theories of consciousness fall on both sides of this line of divide, sparse-illusion views seem to be more common and more influential. I think that views of this kind present two main advantages, which might account for their popularity.

First, given that sparse-illusion views conceive of introspective representations as less rich and detailed than rich-illusion views, they do not have to account for the existence of a heavy *positively inaccurate* introspective machinery. It is a clear advantage, as such heavy and positively inaccurate introspective machinery represents an extra theoretical cost. Indeed, it is not easy to explain *why* we would have such positively inaccurate introspective machinery. For example, how could we give a plausible evolutionary story which would account for the apparition of a complex system of introspective representation which systematically delivers incorrect (and not only schematic) representations?[10] This is particularly tricky given that the illusion of consciousness is supposed to consist in *radically* incorrect representations, that is, in representations of states of certain kinds, which are nowhere to be found in reality. They thus

[8] Note that Humphrey, as Graziano, rejects the term "illusionism" to characterize his theory (Humphrey, 2016).
[9] I am here only classifying recent illusionist views of consciousness. Some more traditional views, which belong to what Keith Frankish calls "weak illusionism" (Frankish, 2016, p. 15), would probably qualify as sparse-illusion views (Armstrong, 1968; Carruthers, 2000; Levin, 2007) – but I will not say more about them here.
[10] This is not to say that all rich-illusion views require an evolutionary explanation of the machinery giving rise to the illusion of consciousness. Given the way I defined cognitive penetrability earlier (a way which exclude diachronic penetrability throughout various stages of development), we can imagine that some views, which would be rich-illusion views in my definition, could explain the illusion of consciousness by appealing to an introspective machinery built or learnt during childhood (for example, on the basis of embodied fallacious cultural schemas or philosophical and religious beliefs). In this case, the illusion of consciousness would not be caused by a species-wide, innate mechanism, which would mean that there might be no need for an evolutionary explanation of the relevant mechanism(s). However, it seems that major contemporary illusionist theories of consciousness which happen to be rich-illusion views often embrace inneism regarding the machinery leading to the illusion of consciousness, which then creates a pressure to give some kind of evolutionary explanation of the corresponding machinery.

stand in contrast with rich illusions which "simply" correspond to *incorrect* representations of magnitudes, magnitudes (length, size, etc.) which nevertheless are to be found in reality, such as the Müller-Lyer illusion, or the moon illusion. Some things in reality have lengths, sizes, colors, though we sometimes represent incorrectly such properties, but according to the illusionist, nothing in reality is ever *phenomenal*. Because the illusion of consciousness putatively consists in these *radically* incorrect introspective representations, proponents of rich-illusion views probably have to admit the existence of a *sui generis* radically erroneous representational machinery. Indeed, if the representations were not *sui generis*, their radically incorrect aspect would be hard to understand. This *sui generis* erroneous representational machinery is arguably costly to explain from an evolutionary perspective.[11]

On the other hand, the idea according to which introspective representations are partial, sparse, incomplete, can be quite easily understood as a consequence of natural selection. For example, if we follow Graziano's line of thought, we can see the incomplete and schematic character of introspective representations as the result of a trade-off between accuracy and processing resources. As for our tendency to commit *mistakes* and to infer that, because our introspective representations of internal states are simple and schematic, our internal states must really be as simple (and must lack any kind of internal complex hidden structure), this could be seen as the unsurprising outcome of some very general – and adaptive – reasoning heuristics.

[11] The difficulty arises whether the representational trait is supposed to be explained as an *adaptation*, or as a by-product of features which are themselves *adaptations*. Note that, by saying that giving such an evolutionary explanation is a challenge, I do not mean that this challenge could not be met. Humphrey's view can be seen as an attempt to give an evolutionary explanation (in that case, as an *adaptation*) of the illusion of consciousness, probably conceived as a rich illusion. In Humphrey's view (Humphrey, 2011), the illusion of entering conscious states, which are represented as eerie, otherworldly states, is adaptive, as it fostered the conviction of our ancestors that human beings and their mental lives had some kind of special value, thus enhancing the drive for survival and reproduction. Of course, one problem with such view is that it seems that the same evolutionary function could have been fulfilled in many different (possibly simpler) ways, without relying on something as complex as the illusory representation of conscious states – for example, by directly designing a strong primitive desire to survive and reproduce. I set aside a more detailed discussion of Humphrey's view, which would go beyond the scope of this paper.

The second advantage of sparse-illusion views is that they do not have to explain why our introspective representations have a content which is such that it positively represents the instantiation of properties which are nowhere to be found in reality. This is a difficult – though probably not unsolvable – problem for illusionism (Frankish, 2016, p. 36‑37). Indeed, if we think about other "standard" *incorrect* representations (other illusions, false beliefs, etc.), their content is usually understood as being grounded either (i) in the representation of properties which are *sometimes instantiated in reality* (e.g. having the illusion that there are two lines of different lengths), or (ii) in the representation of *complex properties* which are never instantiated, while all of their components are *sometimes instantiated in reality* (e.g. having the false beliefs that there is a golden mountain somewhere: while there is no golden mountain, golden things and mountains really exist). But neither (i) nor (ii) seems to be the case with the representation of phenomenal states (Levine, 2001, p. 146-147). Therefore, the manner in which introspection comes to positively represent phenomenal properties, while phenomenal properties are nowhere to be found in reality (nor composed of really instantiated properties), can seem somewhat mysterious. The problem is made even more striking when one considers that many of the main naturalistic semantic theories usually used to account for the content of non-conceptual representations (informational semantics, teleosemantics), of which introspective representations are arguably a species, require that certain *actual* causal links hold between the representations and the represented properties (which in turn seem to require these properties, or their components, to be sometimes instantiated).

On the other hand, sparse-illusion views do not seem to face the exact same problem. Indeed, they do not imply that our introspective representations represent the instantiation of inexistent properties or features, but merely that they represent the instantiation of existent properties and features in an *incomplete* and partial way (which is not particularly problematic to account for). Of course, the representation of inexistent phenomenality has to happen at a

certain level – at the level of the false beliefs that we tend to infer on the basis of our incomplete introspective representations. However, things might get easier here, as it is simpler to see how the content of beliefs (or belief-like states) can be determined by inferential and conceptual role (in a manner described by inferential/conceptual role semantics), so that such content does not require the existence of actual causal links between the representation and the represented properties.

In spite of these two advantages of sparse-illusion views, I think that we should favor rich-illusion views: if consciousness is an illusion, it is a *rich* illusion. I will now give my argument against sparse-illusion views, and in favor of rich-illusion views.

5. An argument against sparse-illusion views

My argument against sparse-illusion views has the structure of a *modus tollens*.[12] It starts with a conditional premise (**premise 1**): if sparse-illusion views are true, then, when we are informed of the fact that our introspective representations of our internal states are in fact schematic and incomplete, so that our internal states have an intrinsic nature which is *not* presented to us through introspection (namely, a physico-functional nature), our reluctance to accept this new piece of information should not be substantively stronger and more resilient than the reluctance we experience when we learn that any of our "standard" incomplete representations really is incomplete. To give examples of the kind of processes which are

[12] The argument I am about to give bears some structural similarity with the argument I gave in (Kammerer, 2018). In both cases, I argue against some illusionist theories of consciousness, on the basis of the fact that they are unable to predict some of our intuitions regarding consciousness. The main differences between the argument here and the argument I gave back then are: (1) My argument back then was directed against *all currently available illusionist theories* of consciousness, while this one is only directed at "sparse-illusion views". (2) My argument back then was based on the inability of illusionist theories of consciousness to correctly predict the strength of *realist intuitions* regarding consciousness (and on the difficulty we face when we try to entertain the idea that consciousness is but an illusion), while I focus here on the strength of *anti-physicalist intuitions* (or "distinctness" intuitions).

precisely supposed to provide a model for the illusion of consciousness according to proponents of sparse-illusion views: we should not be more reluctant to accept the physico-functional nature of our internal states (incompletely represented through introspection) than we are reluctant to accept that the icons on our computers are simply abstract and schematic representations of complex computational structures realized in our computers; or that the *cuteness* of a baby is not some kind of primitive, basic feature of the baby's face, but is its physically-realized dispositional capacity to trigger all kinds of emotional and behavioral reactions in us when we look at it; or that white light is not primitive, simple and pure, but is actually composed of a mixture of all colored lights.

The second premise (**premise 2**) states that we are *more* reluctant to accept the physico-functional nature of our internal states than we are reluctant to accept that any of our "standard" incomplete representations are incomplete. For example, we are more reluctant to accept the physico-functional nature of our internal states than we are reluctant to accept that *cuteness* just is the dispositional capacity to trigger reactions in us, that icons are schematic representations of the complex computational structure of our computers, or that white light is composed of all other colored lights.

The conclusion of the argument is that sparse-illusion accounts are false.

Are the premises of this argument plausible? Let's start with the second premise, which I take to be the least controversial. I think it can be established on intuitive grounds, as well as on sociological grounds. On intuitive grounds first: I may be a convinced physicalist about the human mind, I personally find it incredibly hard to think that what I introspectively grasp when I focus on my stream of experience, really is physico-functional in nature (to the extent that it exists *as it is introspectively presented*). But I am not so reluctant to accept that the icons on my computer's desktop really just are a schematic way to represent a complex structure within my computer. The same goes for the white light case and the cuteness case. In other words: all

of the illusions used by proponents of sparse-illusion accounts are much easier to overcome than the illusion of consciousness. When we *learn* that our representations are incomplete and schematic, and get used to this idea, we manage to no longer infer that there really is something in reality that is *nothing more* than what our schematic representations represent, and we can intuitively accept the true theory about computers (or white light, or cuteness) without much difficulty. This does not seem to be true in the case of consciousness.

This second premise can also be established on sociological grounds. There are still many proponents of anti-physicalism regarding the human mind in general, and consciousness in particular, even amongst educated and clever philosophers. For example, in the Philpapers survey (see http://philpapers.org/surveys and (Bourget & Chalmers, 2013)) taken by 3226 respondents (1803 philosophy faculty members, 829 graduate students), 27,1% of the respondents accepted or leant towards anti-physicalism about the mind (while 56% endorsed physicalism). We can reasonably suppose that a number of them are anti-physicalists about the mind notably because they are anti-physicalists about phenomenal consciousness in particular. Even amongst the philosophers who endorse physicalism, many of them certainly have persistent anti-physicalist intuitions: asked about zombies (creature physically identical to humans, but devoid of conscious experiences), 23,3% of the respondents claimed they are metaphysically possible, and 35,6% took them to be conceivable (though not metaphysically possible). Only 16% of respondents judged zombies to be inconceivable. The conceivability of zombies is used as the first premise of an influential modal argument against physicalism (Chalmers, 1996) ; the wide acceptance of their conceivability can be seen as a rough indirect measure of the wide distribution of anti-physicalist intuitions, even amongst physicalists.[13]

[13] Empirical research on dualist intuitions about the mind has been booming in the last years. See (Chalmers, 2018, p. 13-14) for a recent overview.

On the other hand, virtually no adult who is educated and reasonably smart would maintain that our computers include basic, primitive files corresponding to the icons displayed on the screen; that the cuteness of a baby is a basic, primitive property of the baby (and not the disposition it has to trigger some of our reactions), or that white light is not a mixture of all the colored lights of the spectrum. That shows that there seems to be some peculiar reluctance to accept the fact that all of our internal states that we grasp through introspection in reality have a complex physico-functional nature.

The first premise is maybe more controversial. What is *not* controversial, I think, is a weaker premise: that if the illusion of consciousness is to be conceived of *purely* on the model of user illusions (or similar illusions described by thinkers such as Dennett or Graziano), then indeed we should not be more reluctant to learn about the physical nature of consciousness than we are to learn about the real hidden nature of computer files, cuteness or white light. However, a proponent of sparse-illusion accounts could naturally suggest that these other cases are supposed to be *merely imperfect analogies* of what happens in the case of consciousness. Consequently, we should not think that sparse-illusion accounts are committed to the view that things should happen in the case of consciousness *exactly* as they happen in the case of computer files, cuteness or white light.

This is of course a possible answer. However, in order for the answer to be fully satisfying, the relevant difference between the case of consciousness and these other cases (computer files, cuteness, white light) should be made manifest. Once made manifest, I think that this difference could very well make the resulting theory a rich-illusion view. For example, one could speculate that, in the case of consciousness, the inference from the schematic and incomplete representation to the judgment that there is something in reality that is nothing more than what this representation depicts is *cognitively impenetrable*. That would arguably explain our persistent reluctance to accept the physico-functional nature of our internal states. But that

would also make this account a rich-illusion view. On the other hand, one could stress that our introspective representations of consciousness are not only incomplete (that is, they *do not* represent consciousness as having a physico-functional nature) but that they also feature some kind of meta-content, by which they are *explicitly stating of themselves* that they are complete. For example, these introspective representations of consciousness could represent consciousness as *not having a nature which is not represented in introspection*. This could account for the persistent reluctance we face when we consider physicalism. However, this would again make the view a rich-illusion view, as phenomenal introspection would then consist in the activation of cognitively impenetrable *incorrect* representations – indeed, if illusionism is true, we never enter in states which *do not have* a nature which is *not* represented in introspection.

It might also be possible for sparse-illusion theorists to combine their view with a supplementary independent explanation of our peculiar reluctance to physicalism, without making it a rich-illusion view. In that case, however, it would maybe cease to be a sparse-illusion view. It would perhaps enter the category of "neither-rich-nor-sparse"-illusion view (which I mentioned earlier), given that the explanation of the persistent illusion of consciousness would not be (at least, not only be) that we tend to infer a false belief because of the incomplete nature of our introspective representations of consciousness. For example, it could be that we have an illusory grasp of our conscious states (where the relevant illusion is a *sparse* illusion), but that on top of that the reason why we are particularly reluctant to accept the real physical nature of our internal states is because it contradicts some of our strongly held beliefs – philosophical, religious or intuitive beliefs concerning the immaterial character of the human mind, for example. However, once again, one will be able to ask whether such beliefs are cognitively penetrable or not. If they are, it is unclear how such a view will be ultimately able to account for our *persistent* reluctance to physicalism. If they are not, such an account

amounts to a kind of complex view, which crucially includes elements of a rich-illusion view. Indeed, an account of this kind seems to imply that we are subject to both a sparse *introspective* illusion and a rich *doxastic* illusion about consciousness.[14]

Maybe there are other supplementary hypotheses that the proponents of sparse-illusion views can put forth in order to solve this difficulty (without betraying the spirit of sparse-illusion views), but to the best of my knowledge such hypotheses have not been suggested or discussed in the literature.

Would a rich-illusion view do a better job at solving the problem I just raised for sparse-illusion views? I think so. Let us grant that our introspective representations of phenomenal states positively present them with features which happen to be nowhere instantiated in reality (or positively present them as *not having* some features that they really do have). This would predict that, when we try to think that our internal states are purely physical, we try to think that what we introspect does *not* really have some of the features our introspection represents it as having (or has some features that introspection represent it as lacking). That means that, contrary to what happens according to sparse-illusion views, our learning of the purely physical nature of our internal states does not simply *complete* those of our representations that are generated through a cognitively impenetrable process, but *contradicts* them. But then it is not surprising that we are so reluctant to accept that our mind has this purely physical nature. Indeed, no amount of reflection or of acquisition of new information can make this contradiction *stop*. We are never able to see clearly, no matter how hard we try, how *this* (thought when focused introspectively on our internal states) can be of a physico-functional nature. All we can do is accept that introspection is illusory, that there is nothing that fits what

[14] For an argument that such rich-illusion views in which the rich illusion is a *doxastic* illusion (for example, a cognitive illusion or a fallacy akin to the fallacies studied by psychologists of reasoning (Fisk, 2004; Pohl, 2004; Tversky & Kahneman, 1983; Wason & Johnson-Laird, 1972)) should be rejected, because they are not psychologically plausible, see (Kammerer, 2019a).

is presented to us in introspection, and that our cognitive processes are entirely physico-functional processes.

To go back to the lying metaphor: if we are told lies of omission (and we believe the lies), we can be *surprised* when we learn the full truth, but we can nevertheless maintain our belief in the truth of the incomplete description we were provided earlier. For that reason, our reluctance to accept the new piece of information can disappear on reflection: we can reach that state at which we *understand* how what we were originally told was true, yet incomplete, while we are now in possession of the full truth. On the other hand, if we are told lies of commission (and we believe the lies), we will still be persistently puzzled when we are told the truth. The conflict between the two sources of information will never disappear; we will never reach a state where we *understand* how both can be providing correct representations, but we will simply be forced to discard at least one of them. That shows that rich-illusion views are in a better position to account for the peculiar resistance we encounter when we try to accept the purely physical nature of the human mind.

I tried to show that, in their current form, sparse-illusion views cannot explain the peculiar reluctance we experience when we consider the physico-functional nature of consciousness. This gives us a reason to think that such views are incorrect. On the other hand, rich-illusion views seem to be able to explain that peculiar reluctance. This gives us a reason to prefer views of this kind.

What about "neither-rich-nor-sparse"-illusion views? Even though, as I noted earlier, it seems that most (if not all) of major contemporary illusionist views of consciousness falls either in the rich-illusion or in the sparse-illusion category, the "neither-rich-nor-sparse" category still seems to correspond to an open possibility. I mentioned a possible example of such view earlier – presenting it as a way to "fix" sparse-illusion views, by adding supplementary explanatory factors to explain our persistent reluctance to physicalism (the preexistence of strongly held

anti-physicalist beliefs, of a philosophical, religious, etc., nature). I rejected this view, stating that it faces a dilemma: either the supplementary explanatory factors correspond to cognitively penetrable psychological processes, or they do not. If they do, the corresponding view is unable to account for our persistence reluctance to accept physicalism. If they do not, the view crucially resembles rich-illusion views. I think that other potential "neither-rich-nor-sparse"-illusion views would face the same kind of dilemma. For this reason, I suspect that "neither-rich-nor-sparse"-illusion views are unlikely to provide a genuine alternative to rich-illusion views – although a full assessment of such views would require the direct discussion of putative theories belonging to this category, which do not seem well represented amongst major contemporary illusionist theories of consciousness.

6. Concluding remarks

If we grant that phenomenal consciousness is an illusion, we still have to enquire about the kind of illusion that it is. There are many ways to describe and to specify illusions. One relevant way, I think, is to ask whether it is a rich or a sparse illusion. Many influential proponents of illusionism have developed views which amount to saying that the illusion of phenomenality is a sparse illusion. I think that this kind of view faces a difficulty, as it fails to account for the peculiar reluctance we face when we learn about the real fundamental nature of our mental states. Even though rich-illusion views are less popular, and encounter difficulties of their own (as I stressed in §4), they are in a better position to explain this reluctance. Even views of this kind, of course, have to meet numerous challenges in order to be considered satisfying illusionist theories of consciousness (Chalmers, 2018; Frankish, 2016, p. 29-37; Kammerer, 2018). However, I hope to have shown that rich-illusion views are more likely to deliver the correct explanation of the illusion of consciousness.

References:

Armstrong, D. M. (1968). The headless woman illusion and the defence of materialism. *Analysis*, *29*(2), 48-49.

Bourget, D., & Chalmers, D. (2013). What do philosophers belive? *Philosophical Studies*, *170*(3), 1-36.

Carruthers, P. (2000). *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

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

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

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

Clark, A. (2000). A Case where Access Implies Qualia? *Analysis*, *60*, 30-38.

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

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

Dennett, D. (2016). Illusionism as the Obvious Default Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11-12), 65-72.

Dennett, D. (2017). *From Bacteria to Bach and Back*. Norton & Company.

Drescher, G. (2006). *Good and Real: Demystifying Paradoxes From Physics to Ethics*. Bradford.

Fisk, J. (2004). Conjunction Fallacy. In R. Pohl (Ed.), *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgment and Memory* (p. 23-42). Hove, East Sussex: Psychology Press.

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

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

Graziano, M. (2016). Consciousness Engineered. *Journal of Consciousness Studies*, *23*(11-12), 98-115.

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

Humphrey, N. (2016). Redder than Red: Illusionism or Phenomenal Surrealism? *Journal of Consciousness Studies*, *23*(11-12), 35-55.

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

Kammerer, F. (2018). Can you believe it? Illusionism and the illusion meta-problem. *Philosophical Psychology*, *31*(1), 44-67.

Kammerer, F. (2019a). Does the explanatory gap rest on a fallacy? *Review of Philosophy and Psychology*, *10*, 649-667.

Kammerer, F. (2019b). The illusion of conscious experience. *Synthese*. https://doi.org/10.1007/s11229-018-02071-y

Kammerer, F. (2019c). The Meta-Problem of Consciousness and the Evidential Approach. *Journal of Consciousness Studies*, *26*(9-10), 124-135.

Levin, J. (2007). What is a Phenomenal Concept? In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

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

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

McCauley, R. N., & Henrich, J. (2006). Susceptibility to the Müller-Lyer Illusion, Theory-Neutral Observation, and the Diachronic Penetrability of the Visual Input System. *Philosophical Psychology*, *19*(1), 79-101.

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.

Pohl, R. (2004). Introduction: Cognitive illusions. In R. Pohl (Ed.), *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Hove, East Sussex: Psychology Press.

Pylyshyn, Z. (1984). *Computation and Cognition*. Cambridge (Mass.): MIT Press.

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

Shabasson, D. (ms). *The Incorrect Inference Theory of the Illusion of Phenomenal Consciousness*.

https://www.academia.edu/41257351/The_Incorrect_Inference_Theory_of_the_Illusion_of_Phenomenal_Consciousness

Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, *90*, 293-315.

Wason, P., & Johnson-Laird, P. (1972). *Psychology of Reasoning : Structure and Content*. Cambridge (Mass.): Harvard University Press.