[Penultimate draft – Please cite the published version]

Appearance, reality and the meta-problem of consciousness

Giovanni Merlo

Solving the meta-problem of consciousness requires, among other things, explaining why we are so reluctant to endorse various forms of illusionism about the phenomenal. I will try to tackle this task in two steps. The first consists in clarifying how the concept of consciousness precludes the possibility of any distinction between 'appearance' and 'reality'. The second consists in spelling out our reasons for recognizing the existence of something that satisfies that concept.

1. Chalmers's (2018) meta-problem of consciousness is the problem of explaining why there is such a thing as a hard problem of consciousness: why does it seem so hard to explain the existence of consciousness in physical terms?

We can distinguish two aspects of this problem. One aspect has to do with our reasons for finding consciousness hard to explain: what is it about phenomenal properties that makes it so difficult to accommodate them in the physical world? Why does physicalism about consciousness seem implausible (or, at any rate, less plausible than physicalism about other elements of reality)?

The second aspect concerns our reasons for believing in the existence of consciousness in the first place: why do we take ourselves to instantiate phenomenal properties? Why do various versions of eliminativism and illusionism about consciousness seem implausible (or, at any rate, less plausible than corresponding views about other elements of reality)?

This paper takes up the second aspect of the meta-problem, focusing on views that treat consciousness as an illusory phenomenon. It is often said that part of what makes these views unattractive is the fact that they clash with the intuition that, when it comes consciousness, we cannot sensibly distinguish 'appearance' and 'reality' (§ 2). I agree with this diagnosis, but I will try to improve on it by identifying our *a priori* reasons for endorsing that intuition (§ 3). I will then outline a regress argument directed against views that recognize the legitimacy of those reasons but insist that nothing satisfies

our intuitive concept of consciousness (§ 4 and § 5).[1]

2. *Illusionism* about consciousness is the view that the existence of consciousness is a non-veridical illusion. According to illusionists, consciousness does not really exist, it merely *seems* to exist.[2] Thus, as I look at a red apple (or, for that matter, as I hallucinate a red apple), it only seems to me, introspectively, that I am undergoing an experience involving a rich qualitative content. In fact, there is nothing it is like to see (or seem to see) something red. Similarly, there is nothing it is like to hear a high-pitched voice, to taste chocolate, to touch a smooth surface or to feel in pain.

As its proponents admit, there is something deeply counterintuitive in the idea that consciousness might be treated as an illusory phenomenon. Yet illusionism offers significant theoretical advantages. Like any other version of eliminativism, it liberates us from the need to find a place for consciousness in physical reality. But unlike traditional version of eliminativism, it recognizes that consciousness *seems* to exist. This provides it with more resources to explain the force of certain intuitions, including the anti-physicalist ones. What is the reason for our reluctance, then? What grounds, if any, do we have for resisting illusionism?

In his recent discussion of the 'illusion meta-problem' – the problem of explaining why illusionism is so difficult to accept – Kammerer (2018) suggests an answer to these questions:

> [The] 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. [O]ne of the most immediate definitions of "illusion" is "a fallacious appearance" […]. If we think that there is no distinction between appearance and reality when it comes to phenomenal experience [...], then we will think that there can be no illusion of phenomenal experience. (Kammerer 2018, 50)

According to Kammerer, the reason why illusionism about consciousness is more problematic than other forms of illusionism lies in our intuition that, when it comes to

[1] Since I am interested in our reasons (or rational grounds) for resisting illusionism, the explanation I will offer can be seen as a rationalization of that resistance. For an attempt to offer a psychological explanation, see Kammerer (2019).

[2] Here and in what follows, I will focus on what Chalmers (2018) calls 'strong' illusionism. Frankish (2016) offers a defence of this position.

consciousness, we cannot sensibly distinguish 'appearance' and 'reality' in the way that the notion of an illusion presupposes.

I think Kammerer is right about this, but I also think that his suggestion is only the beginning of a satisfactory answer to the illusion meta-problem. If we really have the intuition that, when it comes to consciousness, we cannot sensibly distinguish 'appearance' and 'reality', the question becomes: what are (or could be) our rational grounds for endorsing that intuition?

One possible answer would be that the very 'appearance' of an episode of consciousness seems enough to guarantee the actual occurrence of that episode. For example, any appearance that one has an experience as of a red dress "seems enough to constitute the experience itself" (Kammerer 2018, 53). Consequently, we cannot posit appearances of consciousness and insist that such appearances are fallacious without contradicting ourselves.

But this line of reasoning can easily be resisted. Certainly, if every illusion involves an 'appearance' and every 'appearance' involves an episode of consciousness, the very idea that consciousness might be an illusion cannot be coherently sustained. But this only shows that understanding the terms 'illusion' and 'appearance' in the way just suggested begs the question against illusionism.

Can we find a way of making sense of our intuitions in this area *without* loading the dice?

3. I think we can make some progress towards solving the illusion meta-problem if we try to articulate our dissatisfaction with illusionism in 'topic-neutral' terms – roughly, "terms that do not mention consciousness (or cognate notions such as qualia, awareness, subjectivity, and so on)" (Chalmers 2018, 16). Specifically, I want to suggest that we should set aside the controversial notion of an 'appearance' and explain our reluctance to endorse illusionism in terms of two topic-neutral epistemological principles.

The first principle connects the notion of an illusion with that of immediate justification:

> [Illusion / Justification Principle] If S is having an illusion that *p*, then S is in a mental state that provides him/her with immediate justification to believe that *p*.

The Illusion / Justification Principle (hereafter, 'IJP') says that part of what it is to be the victim of an illusion is to have

immediate justification to believe its content.[3] To illustrate, suppose you are having an illusion that there's a blue elephant in the room. IJP says that, in such a situation, you are in a mental state that provides you with immediate justification to believe that there's a blue elephant in the room. The justification is 'immediate' in the sense that it doesn't proceed from the rest of your beliefs.[4] We wouldn't call it an *illusion* if you were to believe that there's a blue elephant in the room based on testimony, or by inferring this conclusion from other beliefs.

By itself, IJP does not beg the question against illusionism. An illusionist may well accept that, in having the illusion that we are conscious, we are in a mental state that provides us with immediate justification to believe that we are. However, a problem arises for illusionism when we combine IJP with another principle, namely:

> [Justification / Correctness Principle] If S is in a mental state that provides him/her with immediate justification to form a certain phenomenal belief, then the content of that belief is true.

The Justification / Correctness Principle (hereafter, 'JCP') says that part of what it is to have immediate justification to form a certain phenomenal belief is to be right in forming that belief. To illustrate: if you have immediate justification to form the phenomenal belief 'I am in pain', it is true that you are in pain. As Kripke famously puts it, "to be in the same epistemic situation that would obtain if one had a pain *is* to have a pain" (Kripke 1981, 152).

Note that, just like IJP, JCP does not beg the question against illusionism. JCP does not involve the notion of an 'appearance'. Nor does it imply that any of our phenomenal beliefs is correct. It is true, as it stands, the principle is not formulated in topic-neutral terms. But this problem can easily be fixed by replacing the term 'phenomenal' with whatever topic-neutral term the illusionist will use to demarcate the range of beliefs that her theory aims to explain away as based on the illusion of consciousness.

It is clear, however, that an illusionist cannot accept both IJP and JCP, for the conjunction of the two principles is incompatible with the view that consciousness is a non-veridical illusion. To see this, consider the phenomenal belief 'I am

[3] For present purposes, I understand the term 'illusion' broadly, as covering also cases of hallucination.
[4] Cf. Pryor (2005).

conscious'. By IJP, if I'm the victim of an illusion of consciousness, I have immediate justification to believe that I am conscious. But JCP says that if I have immediate justification to believe that I am conscious, I am indeed conscious. And this contradicts the claim that what I am undergoing is a non-veridical illusion. Illusionists must, therefore, reject one of the two principles – but which?

Suppose they reject IJP, saying that the notion of an illusion should be understood neither in terms of the notion of an 'appearance' nor in terms of the notion of 'immediate justification'. Then it seems they will be hard-pressed to explain what they mean by 'illusion' and in what sense their view is different from more familiar versions of eliminativism, according to which our belief that consciousness exists arises, not from an illusion, but from other beliefs.

Perhaps illusionists could meet this challenge by taking inspiration from certain disjunctivist views on which one can be subject to an illusion that *p* in virtue of being in a state that is 'indiscriminable through reflection from' (Martin 2004, 72) or 'has the same cognitive effects as' (Fish 2009, 94) a veridical perception that *p*. Typically, views of this sort deny that we can provide any positive characterization of the illusory state – in particular, their proponents may deny that illusory states involve the instantiation of phenomenal properties.

However, disjunctivists can make this move only because they characterize illusory states parasitically, by appeal to the phenomenology of non-illusory ones. Their slogan is that the 'bad case' is just like the 'good case', except that in the 'bad case' there is no phenomenal character. Illusionists will have to say something much more radical: that the 'bad case' is just like the 'good case', except that, insofar as phenomenology goes, *there is no such thing as a 'good case'*. The risk is that this strategy will leave them with no clear conception of the very illusion that is supposed to lie at the core of their theory.

The alternative is to retain IJP and reject the JCP. This involves denying that, in the case of phenomenal beliefs, immediate justification guarantees correctness. Some illusionists may find this approach more congenial. However, adopting it will expose them to a kind of different criticism – namely that their view is not really a form of illusionism about *consciousness*.

In the present context, we say that an organism qualifies as having 'consciousness' if and only if it instantiates phenomenal properties. Now, it seems to be part and parcel of our concept of a phenomenal property that such properties do *not* admit of cases where one fails to instantiate the property even if one is immediately justified to believe that one does. The

example of pain illustrates exactly this point: if I have *immediate justification* to believe that I'm in pain, I may be wrong about many things – but not about the fact that I'm in *pain*.

Illusionists may be tempted to dismiss this stipulation as a relic of Cartesian orthodoxy. But I'm not convinced they would be right to do so. JCP does not imply that we enjoy infallible access to our own phenomenal properties. The fact that immediate justification guarantees correctness is compatible with forming all sorts of wrong beliefs about how I feel – I may be irrational, lack the relevant concepts, be misguided by misleading evidence, or be in a condition that impairs my judgmental capacities[5] JCP is simply a constraint on the *content* of phenomenal belief – a consequence of what it is for them to qualify 'phenomenal', i.e. to concern what it is like *for the subject* to be in this or that mental state.

Reflection on this point brings out a fundamental reason for dissatisfaction with the illusionist position. IJP and JCP are not just intuitive; arguably, they are *a priori* principles governing the notions they involve. But if IJP and JCP *a priori*, illusionists are caught between a rock and a hard place. If their view targets *consciousness*, it cannot be a kind of illusionism (at least, not if we take illusions to obey IJP). Conversely, if their view is to be a kind of *illusionism*, it cannot target consciousness (at least, not if by 'consciousness' we mean something that obeys JCP). The idea that consciousness might be an illusion is, at bottom, unstable.

4. The strategy outlined in the last section offers realists a way to address the illusion meta-problem without begging the question against illusionism. But illusionists may be unimpressed.

Perhaps, if IJP and JCP are *a priori*, our concept of consciousness leaves no room for the possibility that what satisfies it might be an illusion. But illusionism is a brand of eliminativism, and the whole point of any eliminativist view is that the concept of consciousness *isn't satisfied by anything*. Thus – it may be said – all that follows from the argument of the last section is that illusionists have misadvertized their view: strictly speaking, they shouldn't say that we have an illusion of consciousness; they should say that we have an illusion of something that closely resembles consciousness, but, unlike consciousness, can be an illusory phenomenon. So, what?

Seen in the light of this complaint, the argument of the

[5] Cf. Wright (2015).

last section lends itself to an unflattering analogy. One may try to prove the existence of God *a priori* by claiming that the concept of God includes the concept of existence – that is the gist of 'ontological' arguments. But we all know the problem with arguments of this sort: no matter how deeply existence is built into God's concept, nothing can guarantee that there will be something, in reality, satisfying that concept. Similarly, no matter how central a principle like JCP is to our concept of consciousness, that concept won't guarantee its own satisfaction. We cannot prove *a priori* the existence of some mental feature that one can correctly self-ascribe whenever one can self-ascribe it with immediate justification.

Realists about consciousness might reply that, with illusionism off the table, the burden of proof falls upon eliminativist to show that the problem should be taken seriously. We have *a priori* reasons to believe that there cannot be non-veridical illusions of consciousness – reasons having to do with how the concept of consciousness interacts with the concepts of illusion and immediate justification. But we also have strong *a posteriori* evidence that consciousness exists: we are introspectively *acquainted* with consciousness, and acquaintance makes the existence of consciousness evident to us.

But eliminativists will not be moved by this reply. To use another unflattering analogy, invoking acquaintance to establish the existence of consciousness is like invoking mystical experiences to establish the existence of God. If one is not convinced of God's existence, one will hardly be willing to regard the experiences in question as veridical. Similarly, if one doesn't believe in the existence of consciousness, one will not share the realist's conviction that we are introspectively acquainted with it.

If this is right, realists need an argument doing for consciousness what 'cosmological' arguments do for God: establishing its existence on a different basis than the *a priori* demands of this or that concept, but also independently of any *a posteriori* evidence that an eliminativist is likely to dismiss as spurious. Can they construct such an argument?

5. I want to argue that a problem confronts any illusionists who deny the existence of consciousness but retain IJP. (This doesn't constitute a refutation of illusionism – illusionists can always reject IJP. But if the considerations I offered in favor of IJP in § 3 are on the right track, rejecting IJP can hardly be regarded as an unproblematic move).

Suppose our illusionist says that, on her view, we are

victims of illusions of pain*, where pain* is a mental feature that exactly resembles pain, except that, unlike pain, it can be the object of non-veridical illusions. If IJP holds, whenever we are victims of illusions of pain*, we must be in some mental state – call it pain** – which provides us with immediate justification to believe that we are in pain*. Thus, the question can be raised: can we ever be victims of non-veridical illusions of pain**?

Suppose the illusionist answers 'no'. Then she needs to explain why what she calls 'pain**' is not identical with what the rest of us call 'pain'. Pain** is a mental state that provides one with immediate justification to believe that one is in pain*, and pain* is a state that exactly resembles pain, expect that it admits of non-veridical illusions. If pain** immediately justifies pain* self-ascriptions and, on top of that, does *not* admit of non-veridical illusions, it seems to have a much better claim to qualify as pain than pain* – indeed, its claim seems to be as good as it can get. So, on this horn of the dilemma, it's unclear that the illusionist position still constitutes a form of eliminativism.

Suppose, instead, that our illusionist answers 'yes', allowing that we can sometimes be victims of non-veridical illusions of pain**. She still accepts IJP, so she will admit that, whenever we have a non-veridical illusion of pain**, we are in some mental state – call it pain*** – that provides us with immediate justification to believe that we are in pain**. Note that pain*** cannot be identical with pain** (otherwise it wouldn't be able to figure in non-veridical illusions of pain**). Nor is pain*** identical with pain* (for pain* must be subspecies of pain** if pain** is to provide one with justification to self-ascribe pain* – so, once again, the identity of pain*** and pain* would leave no room for non-veridical illusions of pain**). Thus, a new question can be raised: can we ever be victims of illusions of pain***?

Here, too, our illusionist can answer only 'yes' or 'no', and neither answer will lead her to a comfortable position. If she answers 'no', her position risks collapsing into a notational variant of realism. If she answers yes, she needs to posit yet another mental property – call it pain**** – about which we can ask whether it admits of non-veridical illusions. An infinite regress threatens.

Interestingly, the same dialectic affects also versions of illusionism that don't involve any explicit commitment to IJP. On Pereboom's (2011) view, for instance, the illusion of consciousness is said to depend on the fact that introspection inaccurately represents phenomenal states via 'phenomenal modes of presentation'. But what about these phenomenal modes of presentation? Pereboom says that they are the object of a

further, higher-order misrepresentation: we represent them as phenomenal even if they are not (2011, 28). Now, either this hierarchy of misrepresentations involving phenomenal modes of presentation goes on *ad infinitum* or it stops with a state that is not so misrepresented. In the first case, we have a regress of (actual or possible) misrepresentations. In the second, we end up with a state that the realist may legitimately identify with consciousness.[6]

The question remains whether the kind of infinite regress we are envisaging should be regarded as vicious. I cannot hope to settle this question here, but I want to offer two reasons to think that accepting this regress would be, at the very least, an unwelcomed result.

Consider an infinite hierarchy of pain*, pain**, pain*** etc. The first point I want to make is that we have no independent ground to believe in such a hierarchy. As Shoemaker puts it,

> No one thinks that in being aware of a sensation or sensory experience, one has yet another sensation or experience that is "of" the first one and constitutes its appearing to one in a particular way. (Shoemaker 1994, 255)

Indeed, we have no clear conception of what it would take to draw a distinction between two elements of the hierarchy. What would it be like, for instance, to be in pain****, but not in pain***? Are there any means by which we could distinguish the two cases?

The second point is that an infinite hierarchy of pain*, pain**, pain*** would seem to involve a regress of explanation. The basic idea behind IJP is that, when we form a false judgment based on a perceptual illusion, something about our mental state at the time of the illusion explains why our error is justified. Very roughly: our wrong judgement that the stick is bent is justified in virtue of our being in a mental state that, in normal conditions, makes it probable that the stick is bent. The mental state involves a 'promise' of truth. Now, if we allow that our having the mental state in question is, itself, something about which we can have non-veridical illusions, and that there are no mental states for which the possibility of illusion is ruled out, it seems that this 'promise' of truth will be always reiterated and never kept. As C. I. Lewis puts it, we will end up with "an indefinite regress of the

[6] Pereboom suggests that, at some point in the hierarchy, the misrepresentations will cease to involve phenomenal modes of presentation (2011, 29, fn 41). But this move turns his view into a form of non-illusionist eliminativism, on which consciousness doesn't really *seem* to us to exist - we only *believe* that it seems to us to exist.

merely probable […] and the probability will fail to be genuine" (Lewis 1946, 186). If we are to avoid the regress, we need to recognize a level where appearances are, not just a *promise* of truth, but a *guarantee* of truth. And that is the level of consciousness.

6. Let me conclude. This paper focused on one aspect of the meta-problem of consciousness: what are our grounds for believing that consciousness exists, and doesn't merely *seem* to exist? I outlined two strategies to answer this question. The first is to identify our *a priori* reasons for ruling out the idea that the 'appearance' and the 'reality' of consciousness might come apart – that was the 'ontological' argument of § 3. The second is to argue that, unless we recognize the existence of mental features whose 'appearance' and 'reality' cannot come apart, we end up with an infinite regress of appearances – that was the 'cosmological' argument of § 5. I suggested that that the first strategy is incomplete without the second: consciousness cannot be an illusion, not only because it doesn't make sense to speak of an illusion of consciousness, but because consciousness is the 'unmoved mover' of immediate empirical justification.

University of Stirling

Acknowledgements

The research for this paper was conducted as part of my work for the project *Knowledge Beyond Natural Science*, funded by the John Templeton Foundation. I'm grateful to my colleagues in the project for many illuminating exchanges on the themes of the paper. Thanks also to Fiona Doherty, François Kammerer and an anonymous referee of this journal for their valuable comments and criticisms.

References

Chalmers, D. (2018). “The meta-problem of consciousness”. *Journal of Consciousness Studies* 25 (9-10): 6-61.

Fish, W. (2009). *Perception, Illusion and Hallucination*. New York: OUP.

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

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

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

Kripke, S. (1981). *Naming and necessity*. Oxford: Blackwells.

Lewis, C. I. (1946). *An Analysis of Knowledge and Valuation*. La Salle: Open Court.

Martin, M. (2004). “The Limits of Self-Awareness.” *Philosophical Studies* 120: 37–89.

Pereboom, D. (2011). *Consciousness and the Prospects of Physicalism*. Oxford: OUP.

Pryor, J. (2005). “There is immediate justification”. In Steup, M. & E. Sosa (eds.), *Contemporary Debates in Epistemology*, Blackwell: 181-202.

Shoemaker, S. (1994). “Self-knowledge and ‘inner sense’”. *Philosophy and Phenomenological Research* 54: 249–314.

Wright, C. (2015). “Self-knowledge: the reality of privileged access”. In Goldberg, S. (ed.), *Externalism, Self-Knowledge and Scepticism*, CUP: 49-74.