This is the author's accepted manuscript of an article published in *Journal of Consciousness Studies*, 23(11–12), 256–89 (2016),
© 2016 Imprint Academic, and reprinted in K. Frankish (ed.), *Illusionism as a Theory of Consciousness* (Imprint Academic, 2017). The published version is available at
https://www.ingentaconnect.com/contentone/imp/jcs/2016/00000023/f0020011/art00020. Version of 18 October 2026.

Not disillusioned: Reply to commentators

Keith Frankish

Abstract: This piece replies to commentators on my target article, 'Illusionism as a Theory of Consciousness' (this issue), building on the arguments offered there. It groups commentators together by their attitude to illusionism, classifying them as advocates, explorers, sceptics, and opponents. It expands on the case for illusionism, refines the position, and responds to objections.

> There is nothing more deceptive than an obvious fact.
>
> (Arthur Conan Doyle, 1892, p. 80)

I am grateful to the commentary authors for their contributions. The aim of this special issue is to give the reader a sense of the potential of illusionism as an approach to consciousness, and the commentators do an excellent job of this, both those who defend the approach and those who challenge it. Each commentary deserves a far more detailed reply than there is space for here, so I shall concentrate on the most salient issues for the overall evaluation of illusionism and focus on points of disagreement rather than agreement. (Thus, if I say relatively little about a piece, this should not be taken to mean that I dismiss it; quite the opposite.) To make this reply a smoother read, I shall group similar commentators together, classifying them as *advocates, explorers, sceptics, and opponents*.

1. Advocates

I begin with a group of commentators who offer further arguments in support of illusionism.

Daniel Dennett provides a characteristically robust statement of the case for illusionism as the default theory of consciousness, arguing that we should thoroughly explore the mundane possibility of illusion before turning to exotic theoretical positions, especially when the latter offer few, if any, empirical predictions. I could not agree more, and if I have been less robust in stating the case for illusionism, it is only for tactical reasons. Of course, opponents will say that the methodological principle to which Dennett appeals is not applicable in this case, since illusionism denies the existence of the very thing to be explained. But this is begging the question, which is precisely whether phenomenality is real or illusory. The explanandum is the thing we call 'conscious experience', where it is an open question whether this involves phenomenality or the illusion of it (compare the

inclusive sense of 'consciousness' defined in section 1.6 of the target article). Illusionists agree that we have a potent intuition that phenomenality is real, but they hold that the rational policy (at least given our current, rudimentary understanding of the neuroscience of consciousness) is not to trust it and to pursue an illusionist research programme. If the programme proves fruitful, then the realist intuition may loosen its grip on us (or we may loosen our grip on it).

Of course, at present illusionists can do little to make it seem plausible that this will happen. They can at best offer vague sketches of how the illusion of phenomenality might be generated, which are easily dismissed. But if truth is our aim, then we should be prepared to put our realist intuition to the test. The widespread reluctance to do this suggests that there may be non-epistemic concerns lurking in the background. Perhaps people worry that ceasing to trust the intuition would erode our sense of self or our sympathy for the suffering of others, and feel that we should hold onto it regardless of its truth. Such worries are, I think, misconceived, but they deserve detailed articulation and assessment. (I shall make some brief remarks later, in responding to Katalin Balog.)

Dennett also urges caution in framing illusionist hypotheses and warns against supposing that there is a clear-cut range of questions and theoretical options that can be identified in advance of detailed empirical work (as some passages in my target article may have suggested). I think these points are wholly salutary.

In his commentary, **Jay Garfield** attacks phenomenal realism, arguing that phenomenal properties would be unknowable, that introspection affords no good evidence for their existence, and that belief in them arises from mistaking properties of external objects for properties of the sensory systems by which we perceive them. In making these arguments he draws on Sellars and Wittgenstein, but he goes on to show that similar ideas have long been present in Buddhist philosophy. In particular, he outlines Vasubandhu's view that it is a misconception to think of experience as having dual subjective and objective aspects — a misconception that yields a doubly distorted view of the causal processes involved.

I am, of course, sympathetic to Garfield's arguments. There are points at which phenomenal realists will want to object (arguing, for example, that zombies do not share the same phenomenal beliefs as us and that we have a special kind of epistemic access to our phenomenal properties), but I shall not discuss these objections here (some relevant points are made in section 3 of the target article). Instead, I shall offer a couple of general observations.

First, a comment on the nature of Garfield's illusionism. If I read him right, Garfield sees phenomenal realism as a purely *cognitive* illusion, which consists in the mistaken belief that our experiences have phenomenal properties. This may be correct, but there might also be a quasi-perceptual element to the illusion. It is possible that we have sensory systems that target aspects of our brain activity, and that these systems play a role in generating the illusion of phenomenality (as proposed in Humphrey 2011, for example). I take it that this possibility is compatible with Garfield's arguments, and indeed with rejection of subject/object duality. Such neuro-senses would be on a par with the other senses, including other body-directed ones such as proprioception. The properties they detect would be inner only in a spatial sense, would not be immediately and infallibly known (and would be known by zombies), and would not be phenomenal in any substantive sense (though they might be *represented* as phenomenal). I don't think we should rule out the possibility that such neuro-senses play a role in consciousness.

Second, a comment on Garfield's discussion of Buddhist philosophy. Garfield has done Western philosophers a tremendous service in introducing them to Buddhist philosophical traditions, which, as he shows, contain much of great contemporary interest and importance (see in particular his 2015). I am not qualified to comment in detail on the points he makes, but the fact that illusionist ideas can be found in ancient Buddhist philosophy is in itself significant. Illusionists are sometimes accused of *scientism* — as if only blind science worship could prompt someone to deny the existence of phenomenal properties. I think this is unfair, and the fact that similar views emerged in a quite different intellectual culture long before the development of modern science helps to rebut it. Vasubandhu's illusionism was the product of a long tradition of metaphysical reflection on the nature of the world and our place in it, and the fact that many Western philosophers find illusionism utterly implausible may say more about their cultural horizons than about the nature of consciousness itself.

Georges Rey devotes his commentary to exploring the nature and origin of the intuition that underlies phenomenal realism. He distinguishes w(eak)-consciousness, which involves implementing various computational processes of attention and internal awareness, and s(trong)-consciousness, which involves meeting some additional, noncomputational condition. (These notions correspond to the two senses of what-it-is-likeness distinguished in section 1.7 of the target article.) We have a powerful intuition that we have s-consciousness, but we have no idea what the extra condition might be nor any independent test for its presence. It is often

assumed that we have rationally compelling introspective grounds for believing in s-consciousness, but Rey questions this. Given the elusiveness of the condition and the known fallibility of introspection, there is scope to doubt that we really have s-conscious states, as opposed to merely having the attitudes and reactions we associate with them. Moreover, Rey notes that we have an equally strong conviction that other people possess s-consciousness, suggesting that our concept of it is sensitive to behavioural factors (perhaps including marks of biological life) as well as to introspective ones. All these points are, I think, very well taken.

Rey goes on to offer a Wittgensteinian diagnosis, according to which talk of 'consciousness' (in the strong sense), like that of 'the sky', has a role within a particular everyday linguistic practice, or 'language game', which cannot be smoothly integrated with science. The concept plays a useful role, reflecting everyday needs, interests, and moral concerns, but we cannot specify its conditions of application, and it does not appear to pick out a well-defined natural phenomenon. (I would add that the fact that we cannot specify its conditions of application means that we cannot be sure that they are *not* wholly functional and behavioural.)

I think Rey's diagnosis is useful and that understanding the nature and function of the concept of s-consciousness will be crucial to developing the illusionist case. Our intuitions here may be put to the test in the not-too-distant future, as we create humanoid robots that have w-consciousness and exhibit a rich variety of human-like behaviour (enabling us to interact with and control them using our existing social skills and knowledge). I suspect such machines will provoke conflicting intuitions. When we interact with them, they will strongly activate our concept of s-consciousness, but when we reflect on how they were made and how they work, we shall have a strong intuition that they lack s-consciousness. This may lead to widespread scrutiny of the concept itself, and perhaps to its revision or replacement.

Rey concludes with some cautionary remarks: some aspects of experience (especially of colour experience) seem deeply resistant to illusionist explanation, and the intuition that s-consciousness is real remains tenacious. It is important that illusionists say these things. They do not claim to have explanations for specific features of conscious experience, or even to see how such explanations will go. They simply claim that the illusionist programme is the most promising one, and that our current intuitions about what can and cannot be explained in illusionist terms may not be reliable. Detailed empirical work may open new theoretical and conceptual options. We may never fully dispel the illusion of s-consciousness; it

may be hardwired into our mechanisms of introspection and social perception, just as some visual illusions are hardwired into our visual systems. But recognizing that that is the case will be a major step forward.

One final point: Rey notes that there is a passage in the target article where I speak of phenomenal properties, conceived as non-existent intentional objects, as being causally potent (Rey, this issue, p. 201, fn. 9). Rey dissociates himself from this view: it is the representations of non-existent intentional objects that are causally efficacious, and talk of the objects themselves having certain effects is merely a convenient shorthand. In fact, I agree with Rey on this; the passage in question was loosely phrased.

Amber Ross highlights some epistemological problems for phenomenal realism. Real properties are independent of our beliefs about them. (Reality, in Philip K. Dick's words, 'is that which, when you stop believing in it, doesn't go away' (Dick, 1995, p. 261) — and, we might add, real properties are ones that don't come to be there just because you think they are.) If phenomenal properties do not exhibit this sort of belief-independence, then the natural conclusion is that they are not real but merely intentional objects of our representations, like the content of a fiction or hallucination. As Ross puts it:

> Any view according to which the subject's beliefs about the character of her conscious experience do play a role in determining the facts of the matter about her conscious experience is a non-realist, illusionist type of view. (Ross, this issue, p. 221)

Yet, as Ross shows in some detail, it is very hard to describe a plausible scenario in which a subject has a false belief about the phenomenal character of an experience they are currently attending to.

Of course, it wouldn't exactly help the realist if we could construct such a scenario; for, as Ross notes, we have a strong intuition that we cannot make this kind of mistake (this issue, p. 219). In this respect, then, illusionism is *better* placed to account for the common-sense view of consciousness than phenomenal realism.

I think the line of attack Ross pursues — questioning the coherence of phenomenal realism — is an important one for the illusionist, and one that was perhaps insufficiently stressed in the target article. It is, of course, a line that Dennett has pressed with considerable force over the years. One example he uses is that of change blindness (Dennett, 2005, chapter 4). People can fail to notice repeated shifts of colour in an image, provided each presentation of the image is separated by a brief masking stimulus. In such cases, the colour shifts must be registered at

some level by the subject's visual system, but do they show up in their visual phenomenology? Dennett argues that realists face a dilemma: If they were to experience change blindness themselves, what would they say about their own phenomenology?[1] If they would say that their phenomenal properties changed without their noticing it, then they must accept that we are not authoritative about our phenomenal properties and that, for all we know, they may change all the time without our noticing. If they say that their phenomenal properties did not shift until they noticed the change in the image (that is, registered it cognitively), then it looks as if our phenomenal properties are simply constructions out of our judgements, as illusionists claim. (And if they say they don't know if their phenomenal properties shifted, then it is unclear what could possibly settle the matter.) The upshot, Dennett concludes, is that the notion of a phenomenal property is simply a mess, a source of nothing but confusion. Ross's contribution illustrates the force of considerations like this, which are, I think, still widely underestimated.

James Tartaglia advocates a surprising position: non-physicalist illusionism. I did not consider the possibility of such a position in the target article, since I was concerned with illusionism as a conservative explanatory strategy, but of course illusionism does not *entail* physicalism. One could be an illusionist about consciousness while holding that reality is fundamentally non-physical. Indeed, Tartaglia argues that illusionism actually provides grounds for holding that.

In the first part of his commentary, Tartaglia attacks 'intermediate' positions, which attempt to combine physicalism with phenomenal realism. Such positions typically rely on the phenomenal concept strategy, but, Tartaglia argues, this reliance is unwise since phenomenal concepts aren't simply neutral ones, which do not present their objects as physical, but substantive ones, which present their objects as having a qualitative, subjective nature that no physical property could have. If physicalism is true, then phenomenal concepts must misrepresent their objects:

> So if the phenomenal concept is a concept of a brain state, it must be a radical misconception of it; we must be misconceiving the brain state beyond all recognition, in fact. We are thinking of a brain state as a subjective experiential array, but that is not what it is at all. Consequently, the array must be an illusion, even if thinking about it somehow allows us to think about real brain states. (Tartaglia, this issue, p. 238)

[1] Dennett uses the term 'qualia', but for consistency I'll put the point in terms of phenomenal properties.

Tartaglia has no sympathy with dualist or panpsychist explanations of consciousness, and he accordingly adopts an illusionist position. This line of argument is, of course, one that I endorse, and Tartaglia's presentation of it is elegant and compelling.

In the rest of his commentary Tartaglia turns to the metaphysical implications of illusionism. He points out that our metaphysics should be compatible with our manifest situation — with how things seem to us, and specifically with the fact that we seem to be confronted with arrays of phenomenal properties. Physicalist illusionists explain our manifest situation by appealing to our judgements and representations: things seem that way because that is how we are inclined to judge them to be. Tartaglia thinks this has profound epistemic consequences. In particular, he argues that it means we cannot be confident in our physical conception of reality, since that conception was created on the basis of illusory experience. He is not suggesting that we should doubt our science. By taking experience as a guide to reality, we have built up a coherent and detailed picture of an objective world — a picture that has eventually led us to the hypothesis that experiences themselves are illusory. But, he argues, we should not forget our starting point; experience itself has a reality that must be accounted for. Hence, we must supplement our scientific picture of reality with a distinctively philosophical account. There must be an independent reality behind our experience, which transcends the objective world in the same way that the objective world transcends the world of a dream. (Tartaglia denies that this independent reality is mental; his view is not a form of idealism or phenomenal realism; this issue, p. 251, fn. 10.)

What should we make of this? Does illusionism require us to reject physicalism? I am unpersuaded. I place myself in the tradition of Quinean naturalism, which (as Tartaglia notes) denies a sharp distinction between philosophy and science and holds that science, broadly construed, provides our best picture of reality. At any rate, I do not see how positing a transcendent reality could shed any further light on the nature of consciousness and subjectivity. (Tartaglia himself says that we can say 'nothing substantive' about independent reality; this issue, p. 250.) Tartaglia worries that physicalist illusionism renders many other aspects of our manifest situation illusory too, including our sense of being spatio-temporally located. Even if this were so (and I'm not sure it is), I do not see it as a reason to reject physicalism. If certain illusions are important to us, we can continue to live by them, treating them as enabling fictions, which do not need metaphysical underpinning.

Moreover, I think Tartaglia overestimates the negative epistemic implications of illusionism. He suggests that it renders our judgements about our experiences 'completely unreliable' (this issue, p. 244). But this is too swift. Illusionism does not claim that our conscious experiences are wholly illusory, only that their apparent *phenomenal aspect* is. I may be correct to judge that I am currently seeing a red post-box (where red is a reflectance property of surfaces), even though I'd be wrong to judge that the experience has a reddish phenomenal feel. Tartaglia asks how we can correlate experiences with worldly properties, but I fail to see the problem. Evolution has set up the correlations, designing our perceptual systems to reliably track worldly properties (perhaps disjunctive, gerrymandered ones), and by doing science we can get a better understanding of the nature of those properties. Tartaglia doubts that we can 'cherry pick' experience for veridical elements, but that is just what the scientific method has enabled us to do. Even if it is not the fundamental reality, the objective world has a complex structure independent of us, and by applying the scientific method we have acquired a powerful grip on that structure. In the process, we have come to question some of the beliefs we started with, but unless we endorse some form of foundationalism, this is not a problem.

In the same vein, illusionists need not deny that phenomenal concepts represent real properties, albeit under distorted guises. Tartaglia finds it implausible that phenomenal concepts represent properties either of brain states or of distal objects, arguing that the misrepresentation involved would be too radical (this issue, pp. 244–5). Again, I fail to see the worry. A concept may track a certain property even if it radically misrepresents it. Recall Humphrey's example of the Penrose triangle, discussed in the target article (Frankish, this issue, p. 17). Deployed in visual experience, the concept of a Penrose triangle tracks a certain sort of three-dimensional structure (a 'Gregundrum'), which it represents as a physically impossible object. The misrepresentation involved is radical yet perfectly possible. Moreover, it may be useful if we need to distinguish Gregundra from non-Gregundra. Gregundra are simply objects that create the illusion of a Penrose triangle, and the best way to tell if an object is a Gregundrum is to see if it creates the illusion — if our visual system misrepresents it as a Penrose triangle. It would be impossible to develop a veridical perceptual concept that reliably distinguishes Gregundra from all the other similar structures that do not create the illusion. Something similar may be the case with phenomenal concepts. They may pick out highly disjunctive physical properties (either of our cognitive systems or of distal

objects) which it is useful for us to track but which are unified only by the fact that they trigger the concept. The fact that they radically misrepresent their objects is no bar to their performing this function.

Perhaps my Quinean sympathies — or my lack of what Tartaglia calls 'historically-informed metaphilosophical self-consciousness' (this issue, p. 252) — are blinding me to Tartaglia's deeper point. At any rate, as far as the science of consciousness goes, he and I are in agreement: we should adopt an illusionist view.

2. Explorers

I turn now to four commentators I have dubbed *explorers*. These use their commentaries to explore ways of developing illusionism — either building theories, responding to objections, or reviewing experimental evidence. Their papers illustrate how illusionism can form the core of a research programme, which can be supplemented and developed in different ways.

François Kammerer addresses the illusion problem — the problem of explaining how the illusion of phenomenality arises. As he notes, the problem has a particularly hard aspect. It is not just that we are strongly disposed to think that phenomenality is not an illusion; we find it hard to understand how it *could* be an illusion:

> what makes us reluctant to accept illusionism is not only that we are disposed to believe that we are conscious, it is also that we have difficulties *making sense of the hypothesis that we are not conscious while it seems to us that we are*. (Kammerer, this issue, p. 127)

This, Kammerer argues, sets this illusion of phenomenality apart from all other illusions and means that it cannot be usefully modelled on them.

Kammerer proposes a solution. Simplified somewhat, it runs as follows.[2] Introspection is informed by an innate and modular theory of mind and epistemology, which states that (a) we acquire perceptual information via mental states — experiences — whose properties determine how the world appears to us, and (b) experiences can be fallacious, a fallacious experience of A being one in which we are mentally affected in the same way as when we have a veridical experience of A, except that A is not present. Given this theory, Kammerer notes, it is incoherent to suppose that we could have a fallacious experience of an experience, E. For that would involve being mentally affected in the same way as when we have a veridical

[2] I have omitted a lot of Kammerer's detail, but I hope I have captured the core of his argument.

experience of E, without E being present. But when we are having a veridical experience of E, we are having E (otherwise the experience wouldn't be veridical). So, if we are mentally affected in the same way as when we are having a veridical experience of E, then we are having E. So E is both present and not present, which is contradictory. (Kammerer couches the argument in terms of experiences, but it could easily be recast in terms of the phenomenal properties of experience. Having a fallacious experience of a phenomenal property involves being mentally affected in the way one would be if the property were present, which involves it being present. Generalized further, this argument might explain our sense that introspection is infallible.)

Kammerer proposes that this explains the peculiar hardness of the illusion problem. The illusionist thesis cannot be coherently articulated using our everyday concept of illusion, which is rooted in our naïve concept of fallacious experience. Moreover, if the naïve theory Kammerer sketches does inform our introspective activity, then we shall not be able to form any imaginative conception of what it would be like for illusionism to be true. Hence the common claim that where consciousness is concerned, appearance is reality. As Kammerer stresses, this does not mean that illusionism actually is incoherent. It simply means that in order to state it we must employ a technical concept of illusion — as, say, a cognitively impenetrable, nonveridical mental representation that is systematically generated in certain circumstances.

Kammerer's approach to the illusion problem is, I think, a promising one, and the idea that introspection is theoretically informed is likely to figure prominently in any developed illusionist theory. Of course, even if Kammerer is right about the source of our intuitive resistance to illusionism, this would not show that illusionism is true, though it would help to dispel one common objection to it. Realists will say that phenomenality is not an illusion even in a technical sense: our relation to our phenomenal properties is one of direct acquaintance, which does not depend on potentially fallible representational processes. Perhaps Kammerer could employ the strategy again here, arguing that our concept of introspective acquaintance is also a theoretical one. At any rate, considerations like this should help to move the debate forward, beyond the simple assertion that illusionism is unintelligible.

Derk Pereboom has done much to establish illusionism as a respectable approach to consciousness, setting out a carefully articulated illusionist theory (the 'qualitative inaccuracy hypothesis') and showing how it can be used to rebut

standard anti-physicalist arguments (see his 2011). In his commentary, he discusses the form an illusionist theory should take, challenging the functionalist view I suggested in the target article. He makes two points. The first concerns what illusionists should say about illusions of phenomenality themselves. Realists will object that illusionists are still committed to phenomenal realism, since there is something it is like to have the illusion of a phenomenal property. In the target article, I suggested that illusionists should deny that phenomenal illusions themselves seem to have phenomenal properties. Pereboom thinks this won't do, and argues that we should instead explain their apparent phenomenality as a further illusion. If introspection misrepresents quasi-phenomenal states as phenomenal, then it can misrepresent our modes of presentation of those states as phenomenal too. This need not create a regress, Pereboom argues, since there is no reason to think that we also represent those higher-order modes of presentation, or at least that we do so under phenomenal modes of presentation.

It is good to have this proposal on the table. I think it is a coherent position, and I agree with Pereboom that the regress objection is not serious. (The point to stress is that mental states seem to possess phenomenal properties only when introspected, and psychological limitations on the introspection process will naturally block the regress.) It is in some ways a puzzling proposal, however. Why should introspection represent modes of presentation as having the same properties as the states they represent? Why should the representation of an experience feel like the experience itself? Moreover, we may not need to posit higher-order introspective processes in order to account for our sense that illusions of phenomenality would themselves have phenomenal properties. Recall Kammerer's proposal about the theory-laden nature of introspection. If Kammerer is right, then when we try to conceive of an introspective illusion we shall conceive of a mental state that incorporates the original experience, with all its (apparent) phenomenal properties. The apparent higher-order feel may simply be an artefact of the innate theory that informs introspection.

Pereboom's second point concerns the nature of quasi-phenomenal properties. Introspection represents these as intrinsic properties rather than functional ones. Illusionism removes the pressure to think of them in this way, allowing us to incorporate them smoothly into a functionalist account of the mind. But, Pereboom argues, illusionism doesn't require us to adopt a functionalist view. We could regard quasi-phenomenal properties as consisting, at least partially, of *absolutely intrinsic aptnesses*, which form the categorical bases for the causal powers of physical

entities. Pereboom argues that this view not only vindicates our commonsense intuition that phenomenal properties are intrinsic causal powers but also gives the illusionist a stronger response to standard anti-physicalist arguments.

Again, it is good to have this view on the table, though personally I find it unpersuasive. Of course, if one thinks that all causal powers are ultimately grounded in absolutely intrinsic aptnesses, then one will think that the powers of quasi-phenomenal properties are too. But I don't think there are *specific* reasons for thinking of consciousness in this way. As Pereboom acknowledges, we do not need to posit absolutely intrinsic properties in order to rebut the anti-physicalist arguments, and our sense that phenomenal properties are intrinsic ones can be explained as a misrepresentation. As illusionists, we do not need the heavy metaphysical machinery of absolutely intrinsic aptnesses in order to explain why conscious experiences seem to be intrinsic causal powers, and employing it would, to my mind, bring illusionism uncomfortably close to a form of Russellian monism.

I turn now to two contributions from scientists. Much scientific work on consciousness has been conducted, wittingly or not, in a quasi-dualistic spirit. Theorists seek to identify the neural processes that produce consciousness, without offering any explanation of *how* they produce it. This isn't surprising if consciousness is conceived in a realist way; it is impossible to gain any explanatory purchase on such a nebulous phenomenon. But illusionism provides a much more tractable target for scientific investigation. To explain consciousness we need to identify and explain the (broadly representational) processes that collectively constitute the illusion of phenomenality. The two commentaries considered next adopt this perspective.

Michael Graziano provides a clear introduction to his *attention schema theory*, according to which consciousness depends on possession of an internal model of one's attentional processes. Graziano's conception of the explanandum for a theory of consciousness is thoroughly illusionist. As he explains:

> Here by 'consciousness' I mean that, in addition to processing information, people report that they have a conscious, subjective experience of at least some of that information. The attention schema theory is a specific explanation for how we make that claim. … It is a theory of how the human machine claims to have consciousness and assigns a high degree of certainty to that conclusion. (Graziano, this issue, p. 98)

The aim is to explain our sense that we are conscious, rather than consciousness itself as a distinct property. This sense arises, Graziano argues, from the fact that

in addition to representing features of the world and of ourselves, we represent our *mental relation* to things via attention. We have an 'attention schema', which models covert attention (the deep processing of selected information), allowing us to monitor and control it. This model does not provide a detailed representation of the mechanisms involved; rather, it represents attention in an abstract, schematic way, as a sort of private *mental possession* of something. As a result, when we introspect our attentional processes we seem to find an inner world where a subjective self has an immediate grasp of the properties of things, leading us to issue reports like this (which Graziano puts into the mouth of a robot equipped with an attention schema):

> 'my mental possession of the apple, the mental possession in-and-of-itself, has no physically describable properties. It's an essence located inside me. … It's my mind taking hold of things — the colour, the shape, the location. My subjective self seizes those things.' (Graziano, this issue, pp. 102–3)

When we talk of consciousness and its features we are reporting the deliverances of our attention schema. Graziano goes on to outline experimental evidence that consciousness is associated with the control functions of the attention schema, as the theory would predict.

I shall not attempt to assess attention schema theory here but simply comment on its relation to the illusionist programme. Graziano notes the affinity (especially with Dennett's views) but argues that the term 'illusionism' has misleading connections. To describe consciousness as an illusion suggests that it is nothing at all and that introspection is simply in error. But, he points out, covert attention itself is real, and our internal model of it, though schematic and abstract, is well adapted for its function of tracking and controlling attention. As he puts it, 'consciousness is not an illusion but a useful caricature of something real and mechanistic' (this issue, p. 112).

I think these are excellent points, and they indicate the need for an important clarification. Talk of illusion does double duty within illusionist theorizing. On the one hand, it may refer to *quasi-perceptual* introspective representations generated by self-monitoring processes, such as the attention schema. These representations may be highly abstract and distorted, and in that sense illusory, but they may also carry valuable information for the system and facilitate important tasks of control and self-manipulation. An illusion need not be a fault and may have

been carefully designed. (Compare Dennett's analogy with the 'user illusions' produced by the icons and pointers on a computer desktop; Dennett, 1991) On the other hand, illusion talk may refer to the *cognitive* illusion involved in judging that we are acquainted with an internal world of intrinsic phenomenal properties. Here it is appropriate to talk of error (certainly in theoretical contexts), though perhaps still not of a fault; belief in the metaphysical specialness of our inner lives may be adaptive, playing an important role in human psychology and social interaction (Humphrey, 2011).

These illusions, quasi-perceptual and cognitive, are of course closely related; we judge that we are acquainted with phenomenal properties because introspection gives us such a partial view of internal reality. (Indeed, natural selection may have sculpted our neural processes in order to create the cognitive illusion; Humphrey, *ibid.*). Phenomenal consciousness, we might say, is a theoretical illusion built on an introspective caricature.

In their commentary, **Nicole Marinsek** and **Michael Gazzaniga** look at illusionism from the perspective of split-brain research. Patients who have undergone surgical severing of the corpus callosum display various behavioural dissociations, which suggest that each hemisphere is operating as a separate mind. This presents a challenge for illusionism. Both hemispheres show signs of being phenomenally conscious (in the everyday sense), so if phenomenality is an introspective illusion, then both must possess a capacity for introspection and be susceptible to illusions. Marinsek and Gazzaniga review relevant experimental evidence and tentatively conclude that this is indeed the case. One moral of this, they suggest, is that, even without callosotomy, phenomenal consciousness may be fragmented, comprising numerous 'modular illusions' with different characteristics.

I think these points are well taken, and, as Marinsek and Gazzaniga note, the split-brain literature will provide a useful testing ground for detailed illusionist proposals. (It would be interesting to explore its implications for attention schema theory and for Humphrey's 'sentition' theory.) Moreover, the suggestion that consciousness may be fragmented is, I think, an important one. One thing the split-brain literature has shown is that our sense of psychological unity can be illusory; despite the dissociations in their behaviour, split-brain patients continue to feel unified, and they unconsciously confabulate to preserve that feeling. Of course, if we conceive of subjecthood in non-psychological terms, as involving direct acquaintance with phenomenal properties, then it is hard to see how we can establish any objective criteria for identifying conscious subjects. But illusionism

provides a much more tractable approach. To be a conscious subject is (putting it very sketchily) to be a system that produces appropriate introspective representations of its own mental activity and uses them to modulate its activity in appropriate ways. In this sense, we may each incorporate multiple conscious or semi-conscious subjects, either modular or partially integrated with each other.

3. Sceptics

This section looks at contributions from four commentators who, although not full-blown opponents of illusionism, express reservations about the position or feel that it is in some way misguided.

Susan Blackmore distinguishes illusionism from a more cautious view, which she calls *delusionism*. Whereas illusionists deny the existence of phenomenal consciousness outright, delusionists hold that we have many mistaken theories about it. Blackmore expresses reservations about illusionism, but she endorses delusionism, arguing that we are wrong to think that there is a stream of consciousness, with rich, unified, and determinate contents, and a persisting self, which observes it. Whenever we introspect, we always find some conscious experience, and this leads us to think that there is a continuous inner stream of such experiences and an inner self waiting to observe them. But these claims, she argues, are baseless; neither neuroscience nor careful introspection offers any way of determining whether conscious experiences are present at times when we are not actively introspecting. Rather, there are just moments of consciousness, temporary constructions bonding thoughts and perceptions to a representation of the self.

This is a valuable piece, which usefully summarizes ideas that Blackmore has defended at length in earlier work. There are many important issues here, but I shall confine myself to commenting on the relation between delusionism and illusionism. Illusionism clearly entails delusionism; if there are no phenomenally conscious experiences, then there is no continuous stream of them either. Could there be a continuous *illusion* of consciousness? It depends on what kind of illusion we are thinking of. If it is a personal-level cognitive one, which occurs when we actively introspect and judge that we are currently having an experience with such-and-such phenomenal properties, then the answer is obviously no. But illusionists might want to say that there is a continuous subpersonal illusion, or something like it, consisting in the production of abstract, quasi-perceptual representations of neural processes, which are used for internal control purposes

and which form the basis for our phenomenal judgements when they occur. (Perhaps Graziano's attention schema theory supposes something like this.)

What about the converse? Does delusionism entail illusionism? Blackmore thinks it does not; she does not endorse illusionism and seems to accept the reality of immediate conscious sensations. *Prima facie* this seems right; there could be moments of phenomenal consciousness without a continuous stream of it. However, I am not sure this position is stable. If there are such moments, then there are properties of one's brain state at those moments that make it phenomenally conscious — physical properties, let us assume. But then it should be possible, at least in principle, to determine whether our brain states have these properties at times when we are not introspecting, and thus to determine whether or not there is a stream of consciousness. If delusionists deny that this is possible, then, it seems, they should deny that there are such properties and accept that phenomenal consciousness does not exist.[3]

Blackmore closes her commentary by suggesting that our delusions of consciousness are malign memes, which we can, with effort, rid ourselves of. I am unsure about this. It may be true that our conceptions of consciousness and the self are culturally shaped, though rooted in the deliverances of real introspective processes. However, they may not be malign; they may play valuable social and psychological roles, as Humphrey has argued (Humphrey, 2011). As we understand more about why we conceptualize our inner lives in the way we do, we should gain more purchase on these questions, perhaps with beneficial practical consequences.

Nicholas Humphrey uses his commentary to question the value of characterizing consciousness in terms of illusion. In the past, Humphrey has proposed an explicitly illusionist theory, according to which conscious experiences reflect internalized expressive responses to stimuli, which interact with incoming sensory signals to generate complex feedback loops. When these loops are internally monitored, Humphrey argued, they appear to possess strange qualitative and temporal properties, creating the illusion of a magical inner world (Humphrey 2011).

In his commentary, however, Humphrey repudiates the label 'illusionist' and insists that his view is better characterized as a realist or 'surrealist' one (though

[3] There are passages in Blackmore's commentary which suggest that her sympathies are more illusionist than she admits. She writes, for example, that neuroscientists 'will never find the neural correlates of an extra added ingredient — "consciousness itself" — for there is no such thing' (this issue, p. 61).

not, he stresses, in any anti-physicalist sense). He offers two reasons for this. One is tactical: to characterize one's view as the claim that phenomenal consciousness is an illusion is to invite people to ignore or ridicule it; it's 'bad politics' (this issue, p. 122). I shall discuss this worry in a moment. Humphrey's other, more substantive, point is that sensations represent something real and important — namely our evaluative responses to stimuli:

> when considering whether sensations are or are not 'real', we must never let go of the fact that sensations do indeed represent *our take* on stimuli impinging on the body. In doing so they represent some of the objective facts about what's happening: the what, where, and when, for example. But, crucially, they also represent how we *evaluate* what's happening, how we *feel* about it. And this is where phenomenal properties come into their own. Sensations represent how we relate to stimulation using, as it were, a paintbox of phenomenal concepts to depict what it's like for us. (Humphrey, this issue, p. 118)

Sensations, he argues, represent two aspects of stimulation: how we are being stimulated (the objective side) and how we respond to the stimulation (the subjective side). Their phenomenal aspect corresponds to the latter; it represents our subjective take on stimulation. And this aspect, Humphrey argues, cannot be illusory or nonveridical:

> How could you … be experiencing a feel that 'doesn't exist'? To be blunt, I think the very notion of this is absurd. When the sensation represents you as feeling a certain way about the stimulation, *that is all there is to it.* The phenomenal feel arises with the representation, and *thereby its existence becomes a fact*. (Humphrey, this issue, p. 119)

There are two ways of reading this. On one, Humphrey is making a point similar to Graziano's: sensations are not mere illusions but representations of something real and important — our evaluative responses to stimuli, what they mean for us. (It is interesting that both Graziano and Humphrey hold that consciousness is based in a dynamic relation rather than passive awareness. Dennett makes a similar point in his commentary.) This reading is compatible with illusionism in my sense. For *phenomenal properties* may still be illusory. It may be that sensation misrepresents our evaluative responses (which are constituted by complex patterns of efferent neural activity) as simple intrinsic phenomenal feels — that it tells us how we feel in the language of phenomenal fictions. Again, this need not imply any

fault in sensation. The distortion may be necessary to achieve the effect; the representation of a huge swathe of neural activity wouldn't have the same impact as a representation of phenomenal pain, just as a pile of sociological reports on parent-child relations wouldn't have the same impact as a performance of *King Lear*. The analogy is with a skilled magician producing astonishing effects, not a desert traveller hallucinating an oasis. (I would add that, *pace* Humphrey, *genuine* error may be possible on the subjective side as well as the objective one. Introspection may sometimes go awry, representing the presence of an internal response that has not in fact been triggered.)

On the other reading, Humphrey is claiming that representations of our evaluative responses to stimuli *create* or *constitute* phenomenal feels as distinct properties. This, of course, is not an illusionist position but a realist one. There are places in the commentary where Humphrey appears to endorse this view, describing phenomenal properties as an 'inherent feature' of brain activity (this issue, p. 120).

On the whole, however, I think the illusionist reading is more accurate. When Humphrey defends his realism, it is the reality of our *relation* to stimuli that he stresses, not the reality of phenomenal feels themselves — which are, after all, usually characterized as non-relational properties (this issue, pp. 119–20). And when Humphrey talks of phenomenal feels 'aris[ing] with their representation' he may mean that they are intentional objects, which are real *for the subject* — part of their represented inner world. I suspect, then, that Humphrey is still an illusionist at heart; introspection uses a 'paintbox of phenomenal concepts' to represent certain internalized responses that are not really phenomenal. This chimes well with his adoption of the term 'surrealism'. Surrealist paintings distort reality, albeit in a creative, expressive way.

Back to the bad politics worry. Should illusionists adopt a less provocative label for their position? 'Illusionism' does have some undesirable connotations. (Any single-word label would have; 'magicism' would be even worse!) And, as I have stressed, we can employ an inclusive concept of consciousness that does not carry a commitment to phenomenal realism and allows us to affirm the reality and significance of consciousness in a natural way. But as a term for a theoretical approach to consciousness, I prefer to stick with 'illusionism' — at least at this stage in the debate, where the ghost of phenomenality has yet to be exorcised from cognitive science. It is all too easy to adopt a conception of what needs to be explained that encourages scientists and philosophers to ask bad questions and to ignore

good ones. In my view, we need to challenge this misconception head on. There's no point mincing words: we don't have phenomenal properties, only representations of them.

Pete Mandik expresses a sympathetic scepticism about illusionism, agreeing that we do not have phenomenal properties but denying that we are under the illusion of having them. He makes two points. First, the term 'phenomenal' (as used in this context) has no clear content. Attempts to define it cycle uninformatively through a series of synonyms, and illusionists won't want to rely on private introspective ostension to explicate the concept. Mandik's second worry concerns the notion of illusion. Illusions are systematic; it is appropriate to talk of illusion only when a certain stimulus or scenario reliably evokes a certain misperception or fallacious judgement in people. In the case of consciousness, the illusion is supposed to be that when we introspect our experiences, they seem to possess anomalous, inexplicable properties. But, Mandik notes, introspection elicits this judgement in few people outside philosophy departments; most would say their experiences seem perfectly mundane and natural. Mandik is tempted to call his position *meta-illusionism* but settles on *qualia quietism*: questions about qualia, or phenomenal properties, are simply not well enough defined to be worth pursuing.

I am sympathetic to Mandik's quietism (which echoes points Daniel Dennett has made), and to some extent I think the difference between us is one of emphasis. However, I think Mandik overstates his case. Take 'phenomenal' first. The concept is typically introduced via a sort of language game (call it 'the phenomenality language game'), which involves a combination of inner ostension (think of how pain feels, coffee smells, etc.),[4] reflection on the appearance/reality distinction (where is the colour of an after image?), thought experiments (imagine inverts and zombies), and scientific knowledge (science tells us that colours are really 'in' us), supplemented with theoretical claims (phenomenal properties are ineffable, intrinsic, radically private, and so on). This game isn't played only by philosophers; many of the moves are widely known, and children spontaneously invent some of them for themselves. I don't claim that the resulting concept is fully coherent, but it is not contentless either, and I think it is meaningful (and important) to deny that it picks out something real.[5] Moreover, I think there is a genuine introspective

[4] Illusionists can engage in inner ostension, though they will take the objects identified to be merely intentional.

[5] I do, however, doubt that there is any content to the weaker, 'diet' conception of qualia sometimes proposed, which is supposedly independent of the phenomenality language game; see my 2012.

basis to the concept. It is sometimes argued that experience is wholly transparent; that we are aware only of aspects of the external world (including our bodies). But it is plausible to think that experience tells us more than this. As several commentators have argued, conscious experience also tells us about our *relation* to worldly properties — how we feel about them (Humphrey, this issue), or how we are attending to them (Graziano, this issue), or what expectations and reactions they evoke in us (Dennett, 2013). Introspection represents these relations under highly abstract, caricatured guises (creating what Dennett calls the *user illusion*), but in doing so it provides a substantive, though misleading, content to the notion of phenomenality.

Second, what about illusion? As I've mentioned, illusion talk does double duty — referring both to quasi-perceptual introspective representations of the sort just mentioned and to the cognitive illusion involved in judging that introspection acquaints us with phenomenal properties (and that these properties are anomalous or magical). The former might be better described as a caricature rather than an illusion, but I think the latter deserves the title. It is true, as Mandik observes, that mere introspection does not elicit these judgements; one needs to have been inducted into the relevant language game. But with that induction (which is common), and a little reflection, people do reliably endorse phenomenal realism and judge that it presents a hard problem. The case is similar to that of the Monty Hall problem, which Mandik cites as an example of a cognitive illusion. In order to fall for the illusion, one needs some prior (albeit imperfect) grasp of the concept of probability.

Finally, a point about quietism. From a tactical point of view, I think, quietism is not the best approach for the qualia irrealist. People easily fall into dualist ways of thinking, which lead them to ask bad questions about consciousness. To counter this, the irrealist needs to provide a robust explanation of why we are susceptible to dualist intuitions and why they seem so compelling. Simply telling them not to talk about qualia won't do. There is a Bob Newhart sketch in which he plays a psychiatrist whose only advice to a neurotic patient is, 'Stop it!', repeated over and over. Quietism is a bit like that. It may be sound advice, but it's not very helpful therapeutically.

Eric Schwitzgebel also considers the problem of defining phenomenal consciousness, responding to the target article's challenge to identify a notion of phenomenal consciousness that is substantive yet free of dubious theoretical commitments. He proceeds by offering a definition by example, describing a range

of uncontentious positive and negative cases and identifying phenomenal consciousness as 'the most folk-psychologically obvious thing or feature that the positive examples possess and that the negative examples lack' (this issue, p. 229). Because it relies on examples and does not adjudicate on contentious cases, this definition is a theoretically innocent one, yet it is not deflationary and leaves room for puzzlement about the nature of consciousness.

I think Schwitzgebel succeeds in identifying an important folk psychological kind — indeed the very one that should be our focus in theorizing about consciousness. However, I don't think he has met the challenge of the target article. For, precisely because his definition is so innocent, it is not incompatible with illusionism. As I stressed in the target article, illusionists do not deny the existence of the mental states we *describe* as phenomenally conscious, nor do they deny that we can introspectively recognize these states when they occur in us. Moreover, they can accept that these states share some unifying feature. But they add that this feature is not possession of phenomenal properties (qualia, what-it-is-likeness, etc.) in the substantive sense created by the phenomenality language game. Rather, it is possession of introspectable properties that dispose us to judge that the states possess phenomenal properties in that substantive sense. (Of course, we could call this feature 'phenomenality' if we want, but I take it that phenomenal realists will not want to do that.) Now, the challenge of the target article was to articulate a concept of phenomenality that is recognizably substantive (and so not compatible with illusionism) yet stripped of all commitments incompatible with physicalism. Schwitzgebel hasn't done this, since his conception is not substantive.

Nevertheless, Schwitzgebel has succeeded in something perhaps more important. He has defined a neutral explanandum for theories of consciousness, which both realists and illusionists can adopt. (I have referred to this as consciousness in an inclusive sense. We might call it simply *consciousness*, or, if we need to distinguish it from other forms, *putative phenomenal consciousness*.) In doing this, Schwitzgebel has performed a valuable service.

4. Opponents

This section responds to four papers by opponents of illusionism. Each makes important points, which deserve more discussion than there is space for here, but I shall indicate the general lines of reply that I favour. As I stressed in the target article, my aim is not to refute alternative positions but simply to establish the

attractions of illusionism. I begin with two commentators who write from a conservative realist perspective.

Katalin Balog mounts a forthright defence of commonsense phenomenal realism. In particular, she argues that explanatory gap considerations do not give physicalists reason to prefer illusionism, since they can be explained by a version of the phenomenal concept strategy. Specifically, Balog proposes that direct phenomenal concepts are partly constituted by the experiences they refer to and refer to them in virtue of this fact; phenomenal states serve as their own modes of presentation. This, she argues, gives us a direct and substantial access to our phenomenal states, which is very different from the access science gives us and creates the impression of an explanatory gap. I indicated my misgivings about this strategy in the target article (see also Tartaglia's commentary in this issue), but I shall add a few more comments here.

First, it is not clear how constitutive self-reference is supposed to work. A concept may be partially constituted by a state without referring to it, and Balog does not explain what further factors are involved in the case of phenomenal concepts. More importantly, at best the account explains why we think phenomenal states could be *non*physical; it does nothing to explain how they could be physical. To see this, it is enough to note that we might have a concept of this kind for a non-phenomenal state, such as a propositional attitude. The concept might represent the state in a substantial and direct way, opening a potential gap between facts about it and the neural facts. (Indeed illusionists might accept that phenomenal concepts represent sensory states in this way.) Yet we might feel no resistance to the idea that the state represented is physical; nothing in our grasp of it need *rule out* its physicality or make it difficult to conceive of. Yet that is just what we feel about phenomenal states; we can't understand how they could be physical, and the phenomenal concept strategy sheds no light on the matter. Moreover, Balog makes it clear that she thinks that experiences possess introspectable phenomenal character when they are not being targeted by phenomenal concepts:

> Thinking about [an experience] and simply having the experience will then share something very substantial, very spectacular: namely the phenomenal character of the experience. (Balog, this issue, p. 45)

But if we have a grasp of phenomenal character that is independent of our phenomenal concepts, then we cannot explain away its puzzling features by reference to those concepts. Simple acquaintance will be sufficient to raise all the familiar questions. What is this 'very substantial, very spectacular property'? How do brain

processes generate it? Why is it not detectable from other perspectives? And how can we be aware of it when we are not representing it to ourselves?

Balog raises other objections to illusionism. It is, she says, 'utterly implausible' (p. 42). It 'flies in the face of one of the most fundamental ways the world presents itself to us' (p. 47) and manifests a misguided — and perhaps dangerous — scientism (p. 42). I am not unsympathetic to Balog's worries, but I think they are unfounded. As we have seen, illusionists do not deny the existence of consciousness in the innocent sense defined by Schwitzgebel; they merely offer a different account of its nature. And here, as Balog puts it, 'the question comes down to the epistemic authority accorded to introspective awareness vs. scientific theorizing' (p. 47). In developing a theory of consciousness, I plump for the latter. We have abundant evidence of the unreliability of introspection, and there is no reason why an evolved cognitive system should represent its internal states to itself in a transparent way, as opposed to an adaptively useful one. I do not accept that this view manifests a scientistic attitude. My motive for adopting it is the same as that for relying on scientific theorizing to explain any other aspect of the natural world — namely, a desire to have an account of the phenomenon that is as far as possible undistorted by human interests and biases. But seeking such an account need not involve dismissing or devaluing other ways of describing the world, including folk theories, the humanities, the creative arts, and spiritual traditions, all of which may pick out patterns and capture insights that are not tractably expressible in the language of science. Nor does endorsing illusionism require us to give up the language of phenomenality; we may continue to employ it as a useful way of characterizing our inner life, while recognizing that it is, in Quine's phrase, an essentially dramatic idiom (Quine, 1960, p. 219). (In comparing qualia to fictions, I am not expressing a negative view of qualia so much as a positive one of fiction.)[6]

Balog also claims that I illicitly appeal to qualitative properties in presenting the case for illusionism. One of her worries concerns introspective phenomenal concepts. If illusionism is true, she points out, these will be either universally misapplied or meaningless. In the former case, it will be miraculous that we have them, and in the latter they will be merely 'mental junk'. It is, she suggests, only because we are already acquainted with phenomenal properties that we can make sense of our having introspective representations that refer to them.

---

[6] Balog suggests that I take a 'negative view' of qualia, as indicated by my use of the term 'embarrassed' in reference to them (Balog, this issue, p. 47). In fact, I used the term to refer to the difficulty *realists* face in accounting for the potency of consciousness — a difficulty illusionists avoid.

I accept that providing a theory of content for phenomenal concepts will be a major challenge for illusionism, but I don't think we have reason at this stage to write it off as unsurmountable. Balog argues that the problem is particularly hard because phenomenal concepts, unlike other nonreferring ones, are simple and direct ones, with no compositional structure. I suspect this is wrong and that the apparent simplicity of phenomenal concepts belies a lot of structure, which we shall tease out as we learn more about the mechanisms involved. (For example, it is plausible that phenomenal concepts contain a distinct affective component. Consider sufferers from pain asymbolia, who recognize pains but no longer find them unpleasant or distressing. Is their introspective concept of pain the same as ours, or a thinner one, which lacks an affective dimension?) As I indicated in the target article, illusionists can appeal to a wide range of factors to explain phenomenal content, including conceptual links, links to non-conceptual sensory and introspective representations, and recognitional capacities for neural states, and they can explain our acquisition of phenomenal concepts as due to developmental processes, individual theorizing, cultural transmission, or a combination of all three.

It may be the case that phenomenal concepts do not pick out metaphysically real (albeit uninstantiated) properties. (I suspect they embody inconsistent theoretical commitments, in which case they presumably do not.) But even if so, it does not follow that they are simply 'mental junk' as Balog puts it. They may still play an important role in expressing our relation to stimuli and orienting us with respect to our own sensory processes.

I grant that it is not easy to see how this approach can account for the apparent richness of our phenomenal worlds, but it is not clear that realists are much better placed to do this. Would the mere existence of a causal connection to a phenomenal property account for the richness of our conception of it? Balog may say that it is our direct acquaintance with phenomenal properties that gives content to our phenomenal concepts. But this is not a genuinely explanatory move, since the acquaintance relation itself is wholly unexplained. (If anything, it is the assumption of acquaintance that is illicit in this context.) My sketch of an illusionist theory of phenomenal content may be, as Balog puts it, hand waving, but it is at least hand waving in the direction of a coherent research programme.

Balog's other worry concerns the functional sense of 'what it is like', which I introduced in contrast to the phenomenal one. To say that one's experiences are *like something* in this functional sense is to say that one has information about

them, provided by functionally defined representational mechanisms. We have a strong intuition that such informational processes would not be sufficient to give us inner lives of the kind we have, but I suggested that if we had a richer and more detailed account of the representations involved, then we might lose this intuition. Balog objects that in suggesting this I am proposing a functional-representational analysis of what-it-is-likeness, which is not only highly implausible but a form of conservative realism rather than illusionism. This mistakes the suggestion, however (which was perhaps not clearly expressed). The idea was not that our *current* notion of what-it-is-likeness is a functional one. I think it is not. Rather, it was that as we develop a richer understanding of the representational processes involved in introspection, we may *reconceptualize* our inner lives in terms of the functional notion of 'what it is like', coming to see them as constituted by representational processes that create the illusion of phenomenality. This wouldn't vindicate the reality of what-it-is-likeness in the phenomenal sense, any more than coming to see a magician's performance as a series of deceptive manipulations would vindicate the reality of the apparent effect.

In his commentary, **Jesse Prinz** argues that illusionism, while not absurd, is less attractive than reductive realism (he dislikes the term 'conservative'), which identifies phenomenal properties with functional or physical ones. He suggests that illusionists, like dualists, expect too much from a physicalist theory of consciousness; identities are not deducible but inferred from correlations and partial explanations, and if a reductive theory can explain enough of the features of consciousness, then we are justified in adopting it.

I take this challenge seriously. Prinz has done tremendously important work in identifying the psychological and neural correlates of consciousness and in providing the sort of partial explanations that he thinks are the best we can hope for in this area (see, in particular, Prinz 2012). But while I do not deny that we may be able to provide reductive accounts of many aspects of conscious experience, I doubt that these will be sufficient to justify realism about phenomenal properties in anything like the traditional sense.

My worries centre on explanatory gaps. While identities may be initially inferred on the basis of partial explanations, we should expect to be able to render them intelligible, giving reductive explanations of higher-level properties in terms of more basic ones. Why should consciousness be an exception, especially when the feature that resists explanation is such a central one? A partial explanationism that doesn't explain phenomenality itself is *too* partial.

Prinz suggests that gaps arise in the case of consciousness because we have direct acquaintance with phenomenal properties. We know our experiences by having them, not by representing them, and this gives us a special sort of knowledge, phenomenal knowledge, which science cannot provide. In the target article I dismissed the notion that physical states can directly reveal themselves to us, but Prinz argues that we can explain it in processing terms. The neural states that constitute conscious experiences directly reveal themselves to us in virtue of being accessible to higher cognition. They do not need to be represented; they are themselves representations and constitute our awareness of external properties. Yet they also tell us something about themselves and the subjective way they present the world to us.

This is an ingenious move, which deserves detailed consideration, but my first response is sceptical. I see how the accessibility of a representational state would give us a grasp of the worldly properties it represents, but I fail to see how it could give us any knowledge of intrinsic properties of the state itself. (Of course, those who adopt a representational theory of consciousness will say that grasping the content of an experience just is grasping its phenomenal character. But then phenomenal knowledge would have no *distinctive* content. I assume Prinz does not want to take this line; indeed, he thinks that phenomenal properties are intrinsic ones; this issue, p. 191.) Prinz argues that conscious states inform us about themselves because they present the world to us in a subjective way, reflecting categories and divisions imposed by our minds, such as categorical colour boundaries. But, arguably, this is just to say that experience *misrepresents* the world in certain ways, and it is unclear how access to a misrepresentation of the world can afford us any knowledge of the representing state's intrinsic nature. Moreover, even if acquaintance did give us knowledge of intrinsic phenomenal properties of neural states, I do not see how this knowledge could have any cognitive significance for us. Neural states affect cognitive processing in virtue of having causal properties that correlate with their representational content. Other causal properties they may possess can have no distinctively *cognitive* effects. So, if phenomenal features are not themselves represented, then they cannot have cognitive effects, even if they have effects of other kinds. This is all very brief of course, but there is a strong case for thinking that all knowledge of the physical world, including those parts of it that constitute our own minds, is representationally mediated.

Prinz also makes some points against illusionism, which he thinks is prone to collapse into reductive realism. His first point concerns the reference of phenomenal terms. We learn these terms, Prinz argues, by pointing to examples, not by description, so if the exemplified states have physical correlates, then it is to these that the terms refer, and realism is true. My response is that a term may be acquired by pointing yet also have a descriptive component. We may learn phenomenal terms from examples but conceptualize their referents as phenomenal in the sense created by the phenomenality language game. (This may involve a later theoretical accretion to concepts originally defined ostensively.) If, as illusionists claim, this conception radically misrepresents the states referred to, then it is misleading to use phenomenal terms for them. Of course, if Prinz is right, then the phenomenal conception does *not* misrepresent those states; but this just takes us back to the issue of whether reductive realism is true.

Prinz's second point concerns the nature of illusions. Illusions require seemings — representations of the illusory situation. What should illusionists say about the seemings that constitute phenomenal illusions? If they say they are phenomenal states, then they have not eliminated phenomenality, but if they say they are beliefs, then they cannot explain the apparent richness of experience, since experience is more fine-grained than belief. I have already discussed this issue in various places, so I shall be brief. I suspect that beliefs may do a lot more work here than we imagine, but the illusionist need not claim that they do it all. Phenomenal illusions may depend on representational states that are fine-grained but not phenomenal. They may, for example, be parasitic on sensory representations themselves (functionally defined), arising when these representations are targeted by conceptual mechanisms. (The thought is that we might introspectively represent a phenomenal property via a sensory content, as the phenomenal property with *this kind* of content.) Or they might depend on quasi-perceptual introspective mechanisms, which give us the sort of caricatured access to our own neural processes that I discussed earlier. These fine-grained representations might be bound up with various beliefs, creating multi-faceted introspective states. This again is hand waving, but I think it is enough to divert *a priori* objections to the illusionist project.

Prinz concludes his comment with some thoughts on the illusion problem. Explaining the illusion of phenomenal consciousness, he suggests, may be no easier than explaining phenomenal consciousness itself and may require very similar resources, in which case the attractions of illusionism diminish. I don't wish to play

down the hardness of the illusion problem, but I think this overstates the case. As a rule, the more magical and inexplicable something seems, the easier it is to create the illusion of it than the reality, so the very hardness of the hard problem should give us reason to think that the illusion problem will be easier to solve. If we resist this conclusion, it may be because of the familiar intuition that there is no appearance/reality distinction for consciousness; the illusion would have to be as magical as the reality. I have already argued that we should not trust this intuition, but I want to add a further point.

Coming to see consciousness as an illusion may involve not only theorizing about the mechanisms involved but also reconceptualizing our inner lives. Compare stage magic again. Working out how a magic trick is done involves resisting our natural interpretation of what we see — our intuitive sense of what forces are at work, what causal sequences occur, what properties items have, and so on. In doing this, we can achieve a sort of aspect switch, reconceptualizing the events we see as a sequence of clever manipulations rather than a simple miraculous effect. In order to understand consciousness, we may need to achieve a similar aspect switch in introspection, reconceptualizing our inner lives as constituted by complex multi-dimensional representational processes rather than simple phenomenal effects. The test will be what happens as we learn more about the psychology and neurophysiology of consciousness and try to apply its findings introspectively. I suspect we shall find that the illusion problem seems increasingly tractable. Of course, we may still default to the intuitive phenomenal aspect, and when we do we may still feel the pull of the hard problem; but this psychological fact will seem increasingly irrelevant.

I turn finally to two commentators who adopt a radical, non-physicalist view of consciousness. **Philip Goff** does not attack illusionism directly but challenges one motivation for it — the claim that radical realism is incompatible with our scientific worldview. The target article offered some reasons for thinking that there is a tension here, but Goff argues that these are not compelling. Radical realism, he argues, is not inconsistent with third-person science, especially as realists can adopt a Russellian monist view, which identifies phenomenal properties with absolutely intrinsic properties of physical entities and systems. (This view is sometimes characterized as a form of physicalism, but it is a nonstandard one, and it is a radical position in my sense.) And although radical realism involves *addi-*

tional metaphysical commitments, beyond those of third-person science, Goff argues that these are not unacceptable. We have introspective grounds for making them, and their disadvantages are not overwhelming.

In reply, I concede that radical realism need not be strictly inconsistent with science, at least if it takes a Russellian monist form. (Interactionism is a different matter, though there is doubtless a lot more to be said on the matter.) If phenomenal properties are absolutely intrinsic ones, then they are simply invisible to third-person science. Moreover, I do not think it is incoherent to suppose that we might supplement our scientific picture of reality in the way Goff proposes, though it might involve some rather extravagant metaphysical commitments.[7] (This isn't to say that I think Russellian monism is without internal problems. In particular, it faces serious problems in explaining how and when subjects of consciousness combine and how subjects correspond to physical organisms.) The case for illusionism is not that there are no alternatives, but simply that it is much *better* than the alternatives — more economical, more elegant, and, most importantly, more explanatory. This is not the place for an assessment of Russellian monism, but I shall say a few words about the last point — explanatory power.

Goff writes as if Russellian monism offers a new framework for research on consciousness, freed from the constraints of what he calls *radical naturalism*. But it is difficult to see how it could do this. Since the theory treats phenomenal properties as intrinsic ones, it offers no predictions as to the behaviour of physical systems (this is what ensures its consistency with third-person science), and its data, which are introspective episodes, cannot be intersubjectively compared and checked. There is no basis here for a collective science of consciousness, but, at best, for multiple individual sciences. Indeed, even this is too optimistic. The data for a radical realist science are *immediate* introspective episodes, and we have no way of comparing such episodes over time or checking that our beliefs about past episodes are accurate. These are not merely 'methodological difficulties', as Goff puts it (this issue, p. 91), but in-principle obstacles to a radical realist science of consciousness. Russellian monism gives us no new explanatory purchase on the world but merely adds a fifth wheel, which serves no function other than to underwrite our conviction that we have direct introspective access to phenomenal properties. Perhaps if all attempts to explain consciousness within the standard scientific framework were to fail, we might fall back on this view. But it would be

[7] Goff himself argues for a *cosmopsychist* form of Russellian monism, according to which the universe itself is conscious (Goff, 2017).

premature to adopt it now. Radical realism may not be incompatible with third-person science, but it is a poor substitute for it.

I shall add one further, minor comment. Goff asks why I assume that consciousness must be known with certainty if it is to be a datum; in general, we needn't be certain of our data (Goff, this issue, p. 92). This is true, but in other contexts we are prepared to question our data in the light of theory. If realists wish to insist that the reality of phenomenal consciousness is a *bedrock* datum, which cannot be questioned (which was the suggestion under consideration in the context of my original remark), then I think they do need to claim that it is known with certainty. Of course, if realists accept that the introspective data are open to question, then it is a different matter. But illusionists will be happy to fight on this ground.

Martine Nida-Rümelin begins her commentary by rejecting the widely held view that phenomenal consciousness consists in having experiences with phenomenal properties. Talk of experiences having phenomenal properties is, she argues, a confused way of talking about *subjects* having *experiential* properties, where these are properties that it is like something to undergo. If this is correct, then it immediately undercuts illusionism as originally presented. If it is confused to think that being phenomenally conscious involves having experiences with phenomenal properties, then it is equally confused to think that it involves having experiences that are misrepresented as having phenomenal properties. However, as Nida-Rümelin notes, illusionists may simply recast their view as the claim that we misrepresent ourselves as having experiential properties, and she goes on to argue against this claim. (In fact, she argues that the claim is *necessarily* false. She accepts that phenomenal consciousness does not fit into our standard scientific worldview and that theoretical considerations would favour illusionism, were it a possible view.) Since this is the crux of the matter, I shall focus on her arguments, granting her earlier move for the sake of argument.

Nida-Rümelin's main argument appeals to facts about reference fixing. She argues that reference to experiential properties should be introduced in the following, two-step way. First, we point to paradigm examples, such as suffering pain, feeling sad, or being visually presented with blueness. Second, we establish reference to a feature all the examples share, using metaphors and provisional descriptions (perhaps talking of 'what it is like' to have the properties), but without making any theoretical commitments as to the nature of the feature. This shared

feature is what marks out experiential properties and thus phenomenal consciousness. Since illusionists deny the existence of experiential properties, Nida-Rümelin argues, they must either deny that this procedure picks out a common feature of experiential properties or say that experiential properties are never instantiated. Neither option, she argues, is attractive.

This argument is similar to Schwitzgebel's, and my response is similar. I grant that the first step picks out real properties — the personal-level properties we call 'being in pain', 'feeling sad', 'seeing a blue colour',[8] and so on. Illusionists do not deny that *something* is going on when we are in pain or feeling sad. And I grant, too, that the second step establishes reference to a common feature of these properties. However, I deny that it is the sort of feature realists think it is. It is not some intrinsic quality, akin to the property characterized by the phenomenality language game. Rather, it is (roughly) the property of having a cluster of introspective representational states and dispositions that create the illusion that one is acquainted with some intrinsic quality. I am sure that this is not what Nida-Rümelin thinks the procedure picks out, but I don't see how she can rule out the possibility. She makes it clear that in the second step reference is to be fixed by ostension, not description (she says that any descriptions used are merely an aid to identification and may not survive later theorizing — this issue, p. 168). So I am happy to concede the truth of realism about experiential properties *in this sense*. However, this is a very weak kind of realism, which is compatible with the ontology of illusionism.

Nida-Rümelin briefly outlines another argument against illusionism, which turns on the claim that our awareness of experiential properties is direct and unmediated:

> If to have a property P and to be aware of having P is one and the same thing, then the awareness of having P cannot possibly 'misrepresent' oneself as having P. On that view, being aware of having an experiential property by having that experiential property does not involve any further step (no reflection, no introspection, no conceptualization) and therefore leaves no room for any kind of illusion. (Nida-Rümelin, this issue, p. 167)

The appeal here is, of course, to direct acquaintance; experiential properties reveal themselves to us immediately, leaving no room for error. This notion has cropped up frequently in this discussion, and it is a central one. *Pace* Prinz, I don't believe

[8] I'm not sure about 'being visually presented with blueness', which is how Nida-Rümelin puts it (this issue, p. 165).

it is possible to give a physicalist account of direct acquaintance in any robust sense. I cannot see how physical properties can reveal themselves in this way to a physical cognitive system. (By 'physical properties' I mean structural and dispositional properties of the sort described by third-person science, not absolutely intrinsic ones.) So as a physicalist, I maintain that our sense of being directly acquainted with experiential properties is itself an illusion, an artefact of the way our sensory and introspective systems are structured. This will not convince Nida-Rümelin, of course, whose metaphysical commitments are different from mine. But I think we are close to bedrock here, and that's a good place to stop.

5. Conclusion

I conclude by thanking the commentators once again. They have forced me to think hard about my position and its commitments, but they haven't shaken my belief in it. If it's an illusion to think that phenomenal consciousness is an illusion, then I'm not disillusioned.[9]

References

Dennett, D. C. (1991) *Consciousness Explained*, New York: Little, Brown.

Dennett, D. C. (2005) *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, Cambridge, MA: MIT Press.

Dennett, D. C. (2013) Expecting ourselves to expect: The Bayesian brain as a projector, *Behavioral and Brain Sciences*, 36 (3), pp. 209–210.

Dick, P. K. (1995) *The Shifting Realities of Philip K. Dick: Selected Literary and Philosophical Writings*, New York: Vintage Books.

Doyle, A. C. (1892) *Adventures of Sherlock Holmes*, New York: Harper and Brothers.

Frankish, K. (2012) Quining diet qualia, *Consciousness and Cognition*, 21 (2), pp. 667–676.

Garfield, J. L. (2015) *Engaging Buddhism: Why it Matters to Philosophy*, Oxford: Oxford University Press.

Goff, P. (2017) *Consciousness and Fundamental Reality*, New York: Oxford University Press.

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

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

[9] I am grateful to Daniel Dennett, Eileen Frankish, Nicholas Humphrey, Maria Kasmirli, and Miloš Tomin for comments, advice, and assistance.

Prinz, J. J. (2012) *The Conscious Brain*, New York: Oxford University Press.

Quine, W. V. O. (1960) *Word and Object*, Cambridge, MA: MIT Press.