CrossMark

ORIGINAL RESEARCH

Against Virtual Selves

Tom McClelland[1]

Abstract According to the virtual self theory (VST), selves are merely virtual entities. On this view, our self-representations do not refer to any concrete object and the self is a merely intentional entity. This contemporary version of the 'no-self' theory is driven by a number of psychological and philosophical considerations indicating that our representations of the self are pervasively inaccurate. I present two problems for VST. First, the case for VST fails to rule out a more moderate position according to which the self exists but is systematically misrepresented by us. This position regards the self as a real entity that has illusory appearances, rather than as a hallucinated entity that has a merely intentional existence. Second, I suggest that this 'illusion model' of self-misrepresentation is preferable to VST. Advocates of VST must acknowledge the existence of an entity—typically the brain—that is the bearer of our misrepresentations of the self. I argue that, other things being equal, we should regard the bearer of our self-representations as the self, even if that entity diverges dramatically from the way we represent the self to be. So by acknowledging the existence of a bearer of self-representations, advocates of VST are in a poor position to deny the existence of the self. I conclude that VST not only fails to rule out the illusion model, but that we have *prima facie* reason to prefer the illusion model to VST.

1 Introduction

According to the virtual self theory (VST), selves are merely virtual entities. On this view, the self is an elaborate fiction generated by the brain. You are *represented* to exist, but these representations do not correspond to any real entity. As such, selves

✉ Tom McClelland
    t.mcclelland@warwick.ac.uk

1   Department of Philosophy, University of Warwick, Coventry CV4 7AL, UK

🖄 Springer

are *merely intentional* entities. VST is a contemporary version of the 'no-self' theory. Although VST differs in various ways from its predecessors, its core thesis remains the same—that the entity you think of as yourself does not exist. In this paper, I will present a two-layered challenge to VST. First, I suggest that the general argument for VST fails to rule out a more moderate position according to which the self does exist but is systematically misrepresented by us. This position regards the self as a real entity that has illusory appearances, rather than as a hallucinated entity that is merely intentional. Second, I suggest that this moderate view is preferable to VST because it avoids certain conceptual worries entailed by denying the existence of selves. These worries revolve around the thought that VST is committed to the existence of a bearer of our representations of the self, and that even if this bearer's self-representations are radically inaccurate a case can nevertheless be made for regarding it as the self. In Sect. 2 I introduce the case for VST. In Sects. 3 and 4 respectively, I argue for the two conclusions above. In Sect. 5 I present some objections and replies and in Sect. 5 I draw some final conclusions.

2 The Case for the Virtual Self Theory

2.1 What is the Virtual Self Theory?

The virtual self theory is not a single well-defined doctrine, but instead a cluster of inter-related positions crossing philosophy, psychology and religion. These positions revolve around the core claim that *selves are merely virtual entities*. This position is characterised by two key commitments: (i) that the brain represents the existence of a self with various properties; (ii) that the entity represented does not exist. This view is articulated in different ways by different authors. Colacurcio proposes that '…our common-sense idea of self as some sort of entity is a human construction, in effect, a virtual reality.' (2012, p. 3) Metzinger claims '…no such things as selves exist in the world: Nobody ever *was* or *had* a self. All that ever existed were conscious self-models…' (2004, p. 1). Bayne similarly states that 'Experiences do indeed have 'owners' or 'bearers', but the owner of an experience is nothing 'over and above' a virtual entity' (2010, p. 290). Finally, Hood claims 'Who we are is a story of our self—a constructed narrative that our brain creates.' (2012, p. xiii).

What does it mean to say that the brain represents the self to exist? Consider the various kinds of mental state we have: perceptual; doxastic; affective; agentive etc. Plausibly, all these mental states are representational.[1] And sometimes, these mental states represent the self: our perceptions of ourselves, beliefs about ourselves, emotions about ourselves and attributions of actions to ourselves are all *self-representations*.[2] These myriad mental representations compose *how our self seems*

---

[1] I am neutral here on whether and how the self shows up in one's phenomenology, though see Farrell and McClelland (2017).

[2] Of course, being a self-representation is not just a matter of representing an entity that happens to be oneself. Self-representations must represent the self *as* the self. This precludes representational states that only contingently represent the self from qualifying as self-representations. It is also worth noting that not

to us. According to VST, this elaborate network of self-representations does not correspond to any real entity: the thing we represent to exist is merely intentional.

Why do advocates of VST claim that the self is a merely *virtual* entity rather than simply saying that selves don't exist? After all, most positions that deny the existence of some class of entity don't describe those entities as 'virtual'. Atheists, for instance, aren't *Virtual God Theorists*. The virtual reality metaphor is integral to VST. Different authors use the metaphor with a different significance, but three main themes can be found in the VST literature.

First, talk of virtual selves acknowledges that we each *appear* to have selves. Indeed, the appearance is so compelling it might be impossible for us to stop representing the self.[3] Hood, for example, proposes that 'Knowing that the self is an illusion cannot stop you thinking it exists…' (2012, p. xvii). For VST, the self is analogous to a convincing virtual reality environment: even though it is not strictly real, the appearance that it is real is vivid and compelling. In other words, the appearance that this non-existent entity is real is *unavoidable*. To appreciate the significance of this, we can again contrast VST with atheism. Most atheists would say that once one has concluded that God does not exist, one can also stop believing that he exists: the appearance that he exists will not persist. Although VST theorists deny that selves exist, they do not think that we can cease representing the self to exist so easily.[4]

Second, VST acknowledges the *utility* of the self-model. No VST theorist thinks the appearance of a self is some kind of unfortunate processing error. Rather, our construction of self-models is a highly-evolved function that serves an invaluable purpose for the organism. Hood, for example, suggests that '[l]ike every other illusion our brain generates, it serves a useful purpose…it provides a focal point to hang experiences together both in the immediate here and now, as well as to join those events over a lifetime.' (2012, p. 290) Metzinger goes as far as saying that '…the conscious self-model is the best invention Mother Nature has made (2004, p. 1). For VST, our self-models are much like a virtual reality environment designed to help us in some real-world task: even though the virtual entities with which we are presented are not real, engaging with this virtual reality may enable us to achieve things in the real world that we would be otherwise unable to achieve. In other words, representing ourselves as having a self is *indispensable*. Again, we can appreciate the significance of this by contrasting VST with atheism. Although most

---

Footnote 2 continued

all self-representation need be *explicit*. Sometimes the self is *implicit* in our representations as a central point around which mental representations revolve (see Bayne 2010; Van Gulick 2014).

[3] Note, even Humeans who deny that we ever come across ourselves *in experience* can agree that there is a way that selves appear to us. Remember, the relevant self-representing states might be doxastic, and even Humeans can countenance the existence of beliefs about the self. Of course, if one goes down the enactivist route of denying the existence of *any* mental representations, then they can deny the existence of self-representations. However, even theorists of this bent will generally agree that there is a way that the self seems to us, even if they don't cash out these seemings in representational terms. Varela et al. (2017) eschew a representational understanding of the mind, but still talk at length about the way the self appears to us and how these appearances do not correspond to any real entity.

[4] This is one respect in which VST differs from certain historical formulations of the no-self view. Varela et al. (2017), for example, explore the Buddhist claim that through disciplined reflection we can free ourselves of the habit of positing an ego.

atheists would acknowledge that belief in God serves some function, they would also claim that we can function perfectly well in the world *without* representing God to exist, and even that we can function *better*. VST theorists think that the value of our self-representations goes deeper than this, and that we could not properly function without our non-veridical representations of the self.

Third, at least some advocates of VST take a distinctive view of the *metaphysics* of selves. Consider the following passage from Bayne:

> Hercule Poirot is a fictional Belgian detective, and his mode of existence is to be contrasted with that of real (non-fictional, actual, existent) Belgian detectives. But there is no kind of real self with which our kinds of selves could be contrasted, for it is in the very nature of selves to be virtual. (2010, p. 27)

Here Bayne is emphasising that selves are virtual entities. A self is constructed from patterns of self-representation, so if the right pattern of self-representation exists then the self exists. These representations lack any concrete independent referent, but to demand such a referent is to misunderstand what kind of thing a self is.[5] Again, the analogy with virtual entities helps capture this metaphysical claim. There are virtual entities, and there are facts about virtual entities. It might, for instance, be true of a virtual environment that it contains a table and false that it contains a chair. A virtual table might not exist independently of the virtual reality mechanism that 'represents' it. But it would be a mistake to say that it does not exist at all, since it is in the nature of virtual tables to have this dependence on a virtual reality mechanism—to be constructed by a system of representations rather than existing as an independent referent of those representations. Once more, the significance of this claim can be drawn out by a comparison with atheism. Atheists do not put God in this subtle representation-dependent metaphysical category. Rather, they flatly deny the existence of God. If one has a network of representations of God, one doesn't thereby construct a real God—one simply has a network of misrepresentations. Although VST theorists are committed to denying that selves exist as independent referents of our self-representations, they typically claim that the self nevertheless has a kind of virtual existence.

The three claims above capture at least part of what it means to regard selves as virtual rather than simply regarding them as non-existent.[6] Although we should be mindful of the significance of the virtual reality metaphor, we should not lose sight of the fact that VST theorists still seek to reject a realist conception of selves that exist independently of our representations of them.

---

[5] A similar line of thought can be found in Dennett (1992) who suggests that to regard the self as anything other than a fictional entity is a category mistake.

[6] When reflecting on the virtual reality metaphor at the heart of VST, we must be careful to avoid a potential misunderstanding. Sometimes 'virtual' entities are understood as a kind of mediator between a subject and a concrete entity. When, for instance, a subject navigates a robot through a real environment using a virtual reconstruction of that environment, they will perceive virtual objects that correspond to real objects. We must be clear that this is *not* the situation proposed by VST. They are not claiming that we have a real self that we can only access through the mediation of a virtual self that stands as its proxy. Rather, they are denying that there is any such thing as a real self behind the virtual entity of which we are aware. They think that the self is a *merely* virtual entity just like the object of a targetless representation is a *merely* intentional entity.

2.2 The Argument for VST

Proponents of VST have offered an eclectic barrage of arguments against the existence of the self. To properly evaluate VST, some order needs to be imposed on these arguments. I suggest that the main arguments offered by advocates of VST follow the same general schema. The schematic argument runs as follows:

(1)   We represent the self as having a particular set of properties.
(2)   No real entity has that set of properties.
(3)   Therefore, the self is not real.
(4)   If the entity we represent is not real, it is merely intentional.
(5)   Therefore, when we represent the self the entity we represent is merely intentional.

Steps 3 and 5 are meant to follow from the premises (though I will seek to cast doubt on this in a moment). Premise 4 is plausibly an analytic truth: the intentional object of a representation is the entity that it represents, but where a representation does not represent any actual thing we say that its object is 'merely intentional' i.e. that it is an intentional object but not a real object. Consequently, the heavy-lifting in the case for VST is done by premises 1 and 2. The case for VST thus revolves around a set of contrasts between how the self appears to us, and how matters stand in reality. My concern is with the validity of the argument that draws on these putative mismatches, not with the mismatches themselves, but it will nevertheless be useful to have some of these putative mismatches to hand. As such, I have summarised the main mismatches proposed in the literature in the following table:

	Appearance	Reality
Endurance	According to our self-representations, there is a single thing that has all of one's experiences across time	According to our best scientific understanding of the brain, there is no such enduring entity. Candidate entities include the organism, the Lockean psychological subject and the stream of consciousness, but none of these are plausibly enduring subjects of experience
Irreducibility	According to our self-representations, the self is an immaterial entity, irreducible to neural processes	According to our best scientific understanding, there are no such irreducible entities and human subjects are wholly material
Free Will	According to our self-representations, the choices made by the self are not pre-determined	According to our best scientific understanding of the world, all events are predetermined (or at least the *probability* of all events is predetermined). Furthermore, according to our best psychological theories, choices can be strongly determined by a combination of environmental and genetic factors

	Appearance	Reality
Agency	According to our self-representations, the conscious choices made by the self initiate our actions	According to our best psychological theories, our actions are initiated prior to our consciously making the choice to perform that act
Central Location	According to our self-representations, the self is located at 'point zero' behind the eyes	According to our best understanding of the brain, there is nothing located at this point that could plausibly be the self
Recollection	According to our self-representations, the self has the history reported in one's biographical memory	According to our best psychological theories, we are vulnerable to severe biases and pervasive confabulation in our memory of ourselves
Essential Character	According to our self-representations, the self has an essential personality that is consistent across times and contexts	According to our best psychological theories, we have no inviolable core of behavioural dispositions, and the way we behave varies hugely across times and contexts depending on the roles we adopt

Although this summary fails to exhaust the philosophical and scientific arguments deployed in the literature, it does serve to show how the case for VST works: one simply plugs an apparent property of the self from the 'Appearance' column into premise 1, and the corresponding observation from the 'Reality' column into Premise 2. This then drives the conclusion that there is no such thing as the self. No advocate of VST cites *all* of these mismatches—each of them can be found in Bayne (2010), Metzinger (2004) or Hood (2012), yet none of these authors cites more than a few of them. But so long as an advocate of VST uses *any* mismatch of this kind in the manner captured by the schematic argument for VST, then the objections I will now introduce should apply.

3 The Case for VST is Unsound

For the purposes of this paper, I will grant both premises 1 and 2 of the argument for VST. This means I am granting: (a) that we do indeed represent ourselves as having a self with the properties listed in the 'Appearance' column and; (b) that the conclusions in the 'Reality' column reveal that no such entity has those properties. I must emphasise that I am only granting this *for the sake of argument*. Although it is plausible that we represent the self to have most of the properties in the 'Appearance' column, I would think that doubts can be raised about whether we really represent the self in some of the ways proposed. Whether our choices even appear to be unpredetermined, for instance, is a contentious matter. Furthermore, although many of the conclusions in the 'Reality' column are well-supported, others are more open to dispute. Regarding agency, for instance, the interpretation of data from the experiments of Libet and his successors is notoriously difficult, and it would be glib to conclude that our actions are definitely not initiated by the choices we make. I would suggest that even a

skeptical reader must acknowledge that at least some of the putative mismatches stand up to scrutiny. But even if one remains skeptical, the point I'm making in this paper is that the case for VST faces a problem *deeper* than that of defending the putative mismatches.[7]

The schematic argument for VST *looks* valid, but closer inspection suggests otherwise. Premise 1 states that we *represent* the self as having certain properties and premise 2 states that no real entity has those properties. The inference that there are therefore no selves is invalid. The problem is that premise 1 is not about how selves must be if they exist. Rather it is about how we represent them to be. Since it is possible for us to *mis*represent real entities, it is possible that there *are* selves but that the way they appear to us is misleading. To see the invalidity of the inference, consider another case of misrepresentation: the perceptual misrepresentation of a straight pencil in water as bent. Here there's a contrast between the properties we *represent* the pencil to have and the properties it *actually* has, yet it would clearly be a mistake to infer that the pencil does not exist. Consider the following argument, which parallels the argument for VST:

(1′) We represent the pencil before us as being bent.
(2′) There is no real entity before us that is bent.
(3′) Therefore the pencil is not real.
(4′) If the entity we represent is not real, it is merely intentional.
(5′) Therefore, when we represent the pencil the entity we represent is merely intentional.

This makes vivid the failing of the schematic argument for VST. Even if proponents of VST are right that our representations of the self are non-veridical, there are two different ways in which we might be misrepresenting matters. First, it could be that the entity we represent is, as VST suggests, non-existent. Second, it could be that the entity we represent does exist, but is not the way we represent it to be, in which case VST is *not* true. Since the case for VST fails to rule out this second possibility, it is an invalid argument.

To get more clear on the distinction between the two interpretations, it will be useful to make an analogy with cases of perceptual misrepresentation. A technical distinction is often made between hallucinations and illusions.[8] In perceptual hallucinations, such as Macbeth's hallucination of a dagger, the subject perceptually represents an object that does not exist. In perceptual illusions, such as the familiar case of the pencil appearing bent in water, the subject perceptually represents an object that *does* exist, but attributes a property to that object (in this case, being bent) that the object does not really have. So, whereas illusions misrepresent their referent, hallucinations fail even to have a referent.

---

[7] More specifically, even if one doubts the premises of an argument it is still worth assessing its validity. I hope to show that even if the putative mismatches do stand up to scrutiny, VST does not follow. This means that how much credence one gives the putative mismatches is by-the-by.

[8] Confusingly, the term 'illusion' is sometimes used in a broad way to encompass both illusions and hallucinations. This is why cases that are strictly hallucinatory are sometimes referred to as illusions.

VST is effectively adopting a *hallucination model* of the self. Johnston eloquently captures the content of such a view: 'Each human being would be a victim of a continuous hallucination, but a hallucination on the subject end of experience, the hallucination of the center' (2012, p. 164, quoted Dainton 2012, p. 182). On this model, our self-representations do not have a real referent. The alternative to this is an *illusion model* of the self according to which self-representations do have a real referent—the self—but systematically and dramatically misrepresent what features the self has. On this model, the self seems to have the properties cited in the 'Appearance' column, but its actual nature conforms to the conclusions in the 'Reality' column.[9] Regarding irreducibility, for instance, the self is a real material entity that presents an illusory appearance of being immaterial.[10] It is this alternative model of the self that is overlooked in the case for VST.

So far I hope to have shown that the premises of the argument for VST only entail the disjunctive conclusion that *either* the self is a merely intentional entity (as per the hallucination model) *or* that the self is a real entity that is misrepresented in our self-representations (as per the illusion model). How significant is the result that VST only entails the disjunctive conclusion? That depends on the plausibility of the second disjunct. If the second disjunct is *implausible* then all the advocate of VST needs to do is supplement their argument to explicitly rule out this unlikely position. If, on the other hand, the second disjunct is plausible then VST would have to revise its dialectical strategy more dramatically to address this alternative proposal. I would suggest that the second disjunct is indeed plausible, and that VST cannot dismiss the illusion model out of hand. To lend support to this verdict, I will consider three quick arguments that advocates of VST might be tempted to employ to dismiss the illusion model and suggest that none of these quick routes is satisfactory.

The first quick argument that might be employed by VST theorists is that our misrepresentation of the self is *too extensive* for the illusion model to be feasible. In the case of the pencil in water, our perceptual representation of the pencil is veridical in many respects: we get the colour of the pencil right, its texture, some aspects of its shape etc. Our illusion is limited to misrepresenting the pencil as bent (and perhaps misrepresenting the submerged half of the pencil as larger than the half above water). This is quite different to the kind of misrepresentation entailed by an illusion model of the self. On that model, we are *radically* wrong about the self. According to our self-representations, the self is an immaterial, undetermined entity with essential behavioural dispositions, but the referent of these representations is really a material entity, bound by determinate laws and lacking in essential behavioural dispositions. But if this entity diverges so dramatically from our self-representations, what sense can be made of the claim that this entity is the referent

---

[9] Advocates of the illusion model include Blackmore (2012), who is careful to specify that when she says the self is an illusion, she is claiming that the self is not as it seems rather than that the self does not exist.

[10] To flesh out the illusion model, it might be necessary to explain *why* we misrepresent the self, but here the illusion model is in much the same position as VST which itself must explain our misrepresentations.

our self-representations? Wouldn't it be more appropriate to say that these representations fail to refer to any real entity at all?

Even if we grant that our misrepresentation of the self is extensive, the problem with this line of argument is that a representation might be of a particular object even in cases where it extensively misrepresents that object. Consider Scrooge's hypothesis that the ghost of Jacob Marley is a figment of his imagination. Scrooge says to the ghost:

> You may be an undigested bit of beef, a blot of mustard, a crumb of cheese, a fragment of underdone potato. There's more of gravy than of grave about you, whatever you are! (Dickens 1843)

Although the apparent object—the ghost of Scrooge's former business partner—has very little in common with an undigested bit of beef, this lack of resemblance needn't count against the plausibility of Scrooge's hypothesis.[11] It is an open possibility that, under the right circumstances, the object of a mental representation differs radically from how it is represented to be. This possibility is exploited in a number of theoretical contexts. Consider Macbeth's dagger again. Although I assumed earlier that Macbeth's dagger was a merely intentional object, others have suggested that Macbeth's experience does have a real object, and so should strictly be regarded as an illusion. Le Morvan (2004), for example, claims that the object of Macbeth's awareness is a real region of air that he radically misrepresents as a floating dagger. Alternatively, Johnston (2004) suggests that the object of Macbeth's awareness is his own brain state—specifically, the state responsible for his visual experience. Both Le Morvan and Johnston have specific theoretical motivations for understanding Macbeth's dagger in this way, but similar example can be found that are unattached to specific theoretical projects.

Although I doubt that the ghost of Jacob Marley was a piece of undigested beef, or that Macbeth's dagger is a region of air or a brain state, the availability of these hypotheses points to something of importance: that there is no a priori obstacle to the claim that the appearance of an object might diverge radically from how it actually is. After all, it would be crude to respond to Le Morvan and Johnston that regions of air and brain states are just too unlike daggers to be the object of Macbeth's dagger-experience. Consequently, emphasising just how inaccurate our self-representations are won't enable advocates of VST to show that those representations have a merely intentional object rather than a real object that they radically misrepresent.[12]

The second argument that VST theorists might deploy to rule out the illusion model draws on the idea that certain properties are essential to an entity being a self. Once one has conceded that there is no entity with those properties, one must accept

---

[11] Here I am wilfully taking Scrooge's utterance at face value. Of course, his words may instead be read as an elision of the claim that a bit of undigested beef is the *cause* of his visual experience but not its object.

[12] Of course, the extent of our misrepresentation is itself open for discussion. If only a few of the putative mismatches between appearance and reality stand up to scrutiny, then perhaps our misrepresentation of the self needn't be so radical. But even if our self-representations are wildly inaccurate, I argue that this would be quite consistent with the illusion model.

that there are no selves. If, for example, it is taken to be essential to selves that they are the enduring bearers of conscious experience across time, this would count against the illusion model—it would render it unintelligible for one to claim that although we represent the self as an enduring bearer of experience it actually is not. By claiming that the object of self-representation lacks a feature essential to something being a self, the advocate of the illusion model is forced to concede that this object is not a self after all.[13]

The problem with this line of argument is that it is very hard to establish what the essential properties of the self are. Even if we grant that the self appears to be the enduring bearer of conscious experience across time, for instance, why should we grant the stronger claim that this is *essential* to something being a self? The same question can be asked of any of the properties in the 'Appearance' column: even granting that the self appears to have that property, why can't it be a property that is only contingent to self-hood? That way, if it transpires that the self does not have that property, this needn't impugn on its self-hood.[14] If the advocate of VST insists that some property is essential to self-hood, there is a worry that this would simply beg the question against the illusion model. According to the illusion model, many of the properties we represent the self to have are misattributed. Perhaps the illusion model can go further and claim we even misrepresent the modal status of those properties, so certain properties seem to be essential to the self that are actually only contingent. Overall, an appeal to essential properties is unlikely to have the dialectical traction needed to rule out the illusion model.

The third quick argument is an application of Occam's Razor. When presented with two competing hypotheses, each equally capable of explaining the data, we should prefer whichever hypothesis is the most parsimonious. Both the illusion model and hallucination model are able to accommodate the mismatches between the appearance of the self and reality. However, the illusion model posits selves where the hallucination model does not. Consequently, the hallucination model is the more parsimonious model so is to be preferred. So although the illusion model is a logically possible explanation of the facts, an inference to the *best* explanation clearly favours the hallucination model.

---

[13] It won't do for advocates of the illusion model to fall back on the claim that self-representations nevertheless have *some* object even though that object is not a self. To see why, consider Macbeth's dagger again: even if Macbeth's experience does strictly have an object—say a region of air, or a brain state—it remains the case that this object won't qualify as a dagger. Just as a case for Macbeth's experience having an object is not a case for the existence floating daggers, so too a case for self-representations having an object is not a case for the existence of selves as such.

[14] Peacocke puts forward a similar line of argument when he responds to Dennett with a distinction between a story about the self and a 'defining' story about the self (2014, p. 72). So long as one's self narrative is not a defining story, it is intelligible to claim that the story is radically false yet the self nevertheless exists. The argument offered by Geddes against Bayne is also similar in spirit. Bayne says that the organism that bears self-representations does not have the features essential to performing the 'self role'. Geddes (2013) responds that these features needn't be regarded as essential, and that the more appropriate response to Bayne's argument is to revise our understanding of the self role in order to retain the idea that the self is identical with the organism.

The problem with this argument is that there is more to explanatory parsimony than minimising the number of entities one posits. When compared under a different frame of reference, it is actually the hallucination model that comes out as more explanatorily profligate. The explanatory task is to account for certain established mismatches between self-representations and reality. The data shows (or so we are granting) that at least some of our representations of the self are non-veridical. The illusion model explains this by claiming that *just these* self-representations are non-veridical because we are subject to particular illusions about the nature of the self. This is quite consistent with the possibility that other self-representations are veridical, including our representations of the self *as existing*. In contrast, the hallucination model explains the particular established mismatches by claiming that *all* self-representations are non-veridical. Since the self does not exist, every one of our self-attributions is false. This suggests that if either model is guilty of going beyond what is justified by the data, it is the hallucination model. Another thing to note here is that the self is not just a theoretical posit. It appears to each one of us that we have a self, and if it appears that *p* then we have defeasible reason to believe that *p*. This suggests that the burden of proof is not on advocates of the illusion model to demonstrate the theoretical need for positing selves. Rather, the burden of proof is on advocates of the hallucination model to demonstrate that every appearance is overturned by some defeating consideration. Although empirical and philosophical considerations might defeat some of the specific appearances of the self—say, the appearance that the self is immaterial—it would be a stretch to claim that they can act as defeaters for every such appearance.[15]

Where does this leave us? Undermining these three quick arguments against the illusion model by no means shows that the illusion model *can't* be ruled out. What it does indicate, though, is that the illusion model cannot be dismissed out of hand. The observation that the argument for VST is consistent not just with the intended hallucination model of the self, but also with an illusion model, is not a mere technical quibble. Rather, it presents a substantive challenge to advocates of VST. Once one has accepted that there is a dramatic mismatch between self-representations and reality, one is presented with two possible views of the self. As things stand, VST fails to give us reason to prefer its hallucination model of the self over an illusion model. This assessment of the limitations of the case for VST is an important conclusion in its own right, and presents a significant challenge to advocates of VST. In the next section, however, I hope to show that it reveals an even more serious problem for VST: that the illusion model is not just a potential competitor to VST but is actually the preferable position.

---

[15] It is worth noting that this response is not just a defence of the illusion model, but actually a positive argument in its favour. The illusion model preserves appearance more than the hallucination model and, other things being equal, we shouldn't posit more error in appearance than is justified by the data. As such, we should prefer the illusion model. Although this is not the argument for the illusion model that I wish to propose in this paper, it is certainly a point in its favour.

4 The Superiority of the Illusion Model

My argument for favouring the illusion model over VST's hallucination model revolves around the following thought: that denying the existence of the self is hard to reconcile with accepting the existence of self-representations, for self-representations must surely have a bearer that can be appropriately described as the self. My development of this argument proceeds in two stages. As a first pass, I suggest an argument against the hallucination model that parallels the *cogito*. The failings of this argument set the stage for an improved second pass that avoids any question-begging assumptions about the self whilst still presenting a significant challenge to VST. Since the illusion model is not affected by these worries about denying the existence of the self, we have reason to prefer it to the hallucination model.

4.1 First Pass: The Hallucination Model as Self-Defeating

It is tempting to say that VST's hallucination model of the self is incoherent. Hallucinations must have a subject, so the claim that the self is itself hallucinated is self-defeating. After all, if I am merely a hallucinated entity, then whose hallucination is it? This line of thought closely parallels Descartes' *cogito*: just as one cannot doubt the existence of the doubter, surely one cannot misrepresent the existence of the representer. A cogito-inspired argument against VST's hallucination model would thus run something like this:

(1) If VST is true, then we each have mental self-representations.
(2) All mental representations have a bearer.
(3) The bearer of one's mental representations is the self.
(4) Therefore if VST is true then there are selves.
(5) Therefore if VST is true then it is false.

It is worth emphasising that VST is indeed committed to the existence of self-representations: VST claims that our self-representations do not refer to a real independent entity, not that those representations don't exist. The argument suggests that accepting the existence of self-representations, or indeed any mental representations, commits one to accepting the existence of a *representer*. Here we see that regarding the self as hallucinatory is quite disanalogous to regarding objects of perceptual experience as hallucinatory. We can make sense of Macbeth hallucinating a dagger because the existence of his dagger-experience is compatible with the non-existence of any actual dagger. In contrast, we cannot make sense of Macbeth hallucinating his own existence because the existence of his self-representations presupposes the existence of a self that does the representing. The argument concludes that VST is committed to a flat contradiction: that there are selves that bear one's self-representations and yet that there are no such things as selves.

If this argument goes through, then we would have good reason to prefer the illusion model of the self to the hallucination model. This is because the illusion model does not deny the existence of selves. According to the illusion model,

although our self-representations radically misrepresent the self, we do still have a real self that is the bearer of these misrepresentations. There is nothing incoherent about the idea of such self-misrepresentation, so the cogito-inspired argument does not apply to the illusion model.

Although there is a certain intuitive appeal to this line of argument, it does not stand up to scrutiny. A familiar objection to the *cogito* is that even if one cannot doubt the existence of a concurrent thought, one can still doubt that one is the thinker of that thought. This powerful objection goes back to Gassendi (1658) and was later echoed by Lichtenberg (1971) and Nietzsche (1886). Nietzsche argued that the existence of a process of thinking does not in itself reveal the existence of an 'ego' doing the thinking. This line of argument was later developed in more detail by Williams (1978). Can a parallel objection be presented to the cogito-inspired argument against VST? The objection would be that even though VST cannot deny the existence of self-representations, it *can* deny that the bearer of these representations is the self. It might be unintelligible to claim that a self-representation exists without *some* bearer, but it is quite intelligible to deny that the bearer of these representations is the self. Dennett captures something close to this line of thought:

> Now how can I make the claim that a self—your own real self, for instance— is rather like a fictional character? Aren't all fictional selves dependent for their creation on the existence of real selves? (Dennett 1992 p. 107)

In order to rebut this intuition, he then goes on to introduce the following analogy:

> I want to imagine something some of you may think incredible: a novel-writing machine. We can suppose it is a product of artificial intelligence research, a computer that has been designed or programmed to write novels. But it has not been designed to write any particular novel…Now imagine that the designers are sitting back, wondering what kind of novel their creation is going to write. They turn the thing on and after a while the high speed printer begins to go clicketyclack and out comes the first sentence. "Call me Gilbert," it says. What follows is the apparent autobiography of some fictional Gilbert. Now Gilbert is a fictional, created self but its creator is no self. Of course there were human designers who designed the machine, but they didn't design Gilbert. Gilbert is a product of a design or invention process in which there aren't any selves at all. (Dennett 1992, p. 107)

Dennett's suggestion is that the brain is analogous to this novel-writing machine. The brain is the originator—the *bearer*—of representations of the self. But it is not the *referent* of those self-representations. Those self-representations refer to a merely intentional entity, not the real concrete entity that generates those representations. This suggests that proponents of VST are free to deny the existence of the self. They require the existence of a bearer of self-representations— whether it be the organism as a whole, the organism's brain, or the brain's narrative-weaving faculties—but they do not require the existence of a self. Generating self-representations is one thing but being the referent of those self-representations is

quite another. The intelligibility of VST's hallucination model can thus be protected against a cogito-style argument.

4.2 Second Pass: The Hallucination Model as Self-Undermining

The first pass at an argument against the hallucination model failed because the existence of mental representations does not guarantee the existence of a self. Nevertheless, a related argument could still tip the balance against VST's hallucination model and in favour of the illusion model. Rather than trying to argue that the hallucination model is incoherent, I will argue for the more modest conclusion that it is inevitably the weaker of the two models available. On this approach, VST's commitments do not *contradict* the thesis that the self does not exist but they do *undermine the plausibility* of such a claim.

Let us start by recalling that both sides posit self-representational states. Although they disagree on whether these states have a referent, they agree that the representations themselves exist (and even agree that these representations are not accurate). Now, if one posits the existence of mental representations one must also posit a bearer of those representations B. To dispense with B whilst leaving its representational states behind would be like dispensing with the Cheshire Cat whilst leaving its smile behind: there is an ontological dependence here that cannot be avoided.[16] Note, the concept of a bearer of representations is not the same as the concept of a self. As such, it is coherent for advocates of VST to say that there exists a bearer of self-representations whilst denying that this entity is the referent of those representations. That said, it is also coherent for the illusion model to identify the bearer of self-representations with the self. I would suggest, moreover, that by dint of being the bearer of self-representations, we have defeasible reason to identify B with the self, and thus to choose the illusion model in favour of the hallucination model. I will now unpack this line of thought in more detail.

The view that $B = S$ is quite compatible with the self being radically misrepresented. Against Dennett, I am suggesting that B is not so much like a novel-writing machine spinning a fiction about a merely intentional entity, but instead like an unreliable diarist spinning an inaccurate narrative about itself. Just as the inaccuracies of the diarist don't change the fact that the author of the diary is identical with its subject, nor do the inaccuracies of our self-representations change the fact that the generator or bearer of those representations is identical with the self. If this line of argument goes through, it would put the burden of proof on the hallucination model to demonstrate that B's self-representations refer to a merely intentional entity rather than representing (however inaccurately) B itself. But why think that the illusion model's claim that $B = S$ is the default view rather than the hallucination model's claim that S is a merely intentional entity?

---

[16] Here the VST theorist may dig in their heels and insist on a bearer-free conception of representations as Humean *happenings* that are not attached to any particular object. I would suggest that such a position is unintelligible. A mental state is a representation only if it represents something *for* a representer. I'm not sure what sense can be made of a representation that *just represents* without having a bearer.

My first answer to this question appeals to our ordinary conception of self-hood. Our conception of self-hood is nebulous and contentious, and uncovering a neat list of necessary and sufficient conditions of self-hood is unlikely. But even if we cautiously avoid the assumption that being the bearer of mental representations is *sufficient* for being a self, we can propose the more cautious principle that being the bearer of mental representations gives us *defeasible reason* to regard something as a self. After all, a whole host of platitudes about the self revolve around it being the bearer of mental states. To say that the self is the agent of one's actions, the thinker of one's thoughts or the subject of one's perceptual experiences is to say that the self is the bearer of those various categories of mental state. One might even make a case for saying that being the bearer of mental representations is at *the very core* of our understanding of what it is be a self. As such, if *B* is the bearer of one's mental representations then we have good reason to regard *B* as the self.

The difficulty with this line of argument is that appeals to our ordinary conception of the self can cut both ways. We understand the self to be the bearer of our mental representations, so to that extent *B* fits the template of self-hood. However, we also understand the self to have the properties listed in the 'Appearance' column of the table above yet, as the illusion model itself concedes, neither *B* nor any other candidate for self-hood has these properties. I have used the fact that *B* meets a certain key criterion of self-hood as a reason to regard *B* as the self, but an advocate of VST can use the fact that *B fails* to meet certain criteria of self-hood as a reason not to regard *B* as the self, and indeed can use those criteria to conclude that no actual entity should be regarded as a self. Perhaps this line of argument can be resisted by claiming that some criteria are more important than others: if being the bearer of mental representations is a *core* criterion of self-hood, perhaps that gives us a reason to regard B as the self that trumps the other considerations raised. However, although I find it plausible that bearing mental representations is fundamental to our conception of self-hood, it is doubtful that a non-question-begging case could be made for such a conclusion. As such, we must adopt a different tack to show that the illusion model is preferable to the hallucination model.

Rather than appealing to our ordinary conception of self-hood, it might be better to appeal to the content of our self-representations. Much is made of the fact that our self-representations at least purport to be *de se* representations. *De se* representations are representations that refer essentially to the representer of that representation. Consider the thought 'my pants are on fire' (Lewis 1979). This is not a thought that just happens to refer to oneself: rather, the first-person mode of presentation of the thought *guarantees* that the referent of the thought is you rather than someone else. Crucially, the thought that self-representations are *de se* representations is not limited to realists about the self. Bayne, for instance, places significant weight on the notion of *de se* representation. He explains: 'In *de se* representation the subject represents themselves as themselves. *De se* representation isn't the exclusive provenance of explicitly self-conscious thought, but permeates consciousness through and through.' (2010, p. 289)

Now, given that self-representations purport to be *de se* representations we are left with the following situation: B is the bearer of self-representations and those self-representations purport to refer to whatever entity is the bearer of those representations. This means that according to our self-representations, whatever entity is the bearer of those very representations is the self. As such, according to our self-representations $B = S$. If it appears to us that $B = S$ then we have defeasible reason to believe that $B = S$. So the burden of proof is on VST to defeat this appearance and give us reason to believe that the referent of our self-representations is not B.

This argument acknowledges the possibility that B is not the referent of our purportedly *de se* representations. One might hold that B generates a fiction in which there is an entity S that is the bearer of representations that refer essentially to S. Alternatively, one might think that although self-representations purport to be *de se* they actually are not: our self-representations present themselves to us as referring to the bearer of that very representation, but rather than referring to B they refer to the fictional entity S. I am not arguing that these possibilities are unintelligible. I am arguing, however, that the burden of proof is on VST to show why we should think that our self-representations are misleading in this way. Although the data cited by VST shows that we misrepresent the self in various ways, it does not give us any direct reason to think that the apparent *de se* character of our self-representations is misleading. In the absence of such data, our default stance should be that our self-representations do indeed refer to the bearer of those representations.

This way of appealing to *de se* representation to challenge VST should not be confused with Van Gulick's argument against Bayne's version of VST. Van Gulick argues as follows:

> the subject or self to which those experiences are *de se* referred…is a merely intentional entity defined as the point-of-view of a set of experiences. However, such a merely intentional entity could not be the self or subject that actually has the experiences. The virtual self is an abstract entity, and as such how could it be the subject that actually has real experiences? It seems it cannot. But if the virtual self is not the same as the subject that actually has the experiences, then how could those experiences refer to that virtual self *in a de se way*…? (2014, p. 497)

Van Gulick is arguing that real representations must have a real bearer, and that real *de se* representations refer to whatever the real bearer of those representations is. This an argument against the very intelligibility of *de se* representation for VST. The limitation of Van Gulick's argument is that it does not acknowledge the possibility that our *de se* representations are a kind of virtual representation—that the representations themselves are part of the fiction generated by B. Nor does he acknowledge the possibility that our purportedly *de se* representations are not actually *de se* representations at all. These radical options are available to VST, and cannot be ruled out a priori. One of the respects in which my argument differs from Van Gulick's is that it uses an appeal to *de se* representation not to try and render

VST unintelligible but rather to shift the burden of proof from the illusion model onto VST.[17]

5 Objections and Replies

5.1 The Straw Man Objection

5.1.1 Objection

The availability of the illusion model is wielded as an objection to VST, but looking at what advocates of VST actually say about the self they seem to be advocating an illusion model rather than an hallucination model. This means that the objections raised in this paper are targeted against an artificially radical formulation of VST, and that there is no real disagreement with the views actually articulated in the literature. Some VST theorists talk of *our* self-representations and of the way the self appears to *me*. For instance, when Metzinger says '…you constantly *confuse* yourself with the content of the self-model…' (2004, p. 1) he seems to be assuming that you do indeed exist and are the bearer of certain misrepresentations of the self. Passages like these are commonplace in the VST literature, and suggest that advocates of VST are only denying the existence of the-self-as-it-appears-to-us, not of the self-as-such. Furthermore, some theorists such as Hood (2012) consistently describe the self as an *illusion* rather than as a *hallucination*. This indicates that they are already adopting an illusion model of the self rather than a hallucination model. The hallucination model is thus a straw man.

5.1.2 Reply

The wording used by advocates of VST can indeed give the appearance that they are not denying the existence of a real self independent of our representations. Careful examination, however, suggests that such wording is misleading and that their real position is indeed that the self does not exist. One can find flat statements that the self is a merely intentional entity. For instance, on the same page as the passage from Metzinger quoted above, he states that '…no such things as selves exist…' (2004, p. 1). Why, then, do we find misleading passages in which these authors seem to attribute self-representations to a subject? The simple answer is that it is built into our language that mental representations belong to a *subject* rather than to some object that falls short of self-hood. It sounds unnatural to talk of 'the brain's misrepresentation of the self', so it is to be expected that advocates of VST

---

[17] Another possible reason for preferring an illusion model to an hallucination is provided by Peacocke, who proposes that self-representations must have a referent for 'It is at the level of reference that the notions of correctness and incorrectness get a grip.' (2014, p. 72) On this argument, VST's claim that our self-representations are untrue *presupposes* that those representations have a referent. Although I agree with the spirit of Peacocke's argument, I worry that advocates of VST can respond that they understand self-representations as something akin to definite descriptions that lack a referent but can still be evaluated as false (at least on Russell's conception).

sometimes fall into a less rigorous way of speaking. Furthermore, when theorists like Hood (2012) talk of the self *illusion* they are not using the term in the technical sense intended in this paper. As noted above, sometimes 'illusion' is used in a way that encompasses misrepresentations that don't have a real object and misrepresentations that do. Hood is clearly using it in this broad sense as he is quite explicit that selves do not exist.

5.2 Indeterminacy of Reference

5.2.1 Objection

The hallucination/illusion distinction seems quite tidy, but there are some cases that reveal the distinction not to be clear cut. There are cases of misrepresentation that could be construed as attributing properties to a merely intentional entity, or could be construed as misattributing properties to some real entity. For instance, hallucinogenic drugs might make one perceptually experience a demon at the window of one's room. It is very hard to determine whether this is an *hallucination* of an entity—the demon—that is not real, or an *illusion* in which a real entity—the curtains—is misrepresented as a demon. It could even be argued that there are cases in which there is *no fact of the matter* about whether some mental representation has an object or not, and thus no fact of the matter about whether it is a hallucination or an illusion. Rosenthal, for example, suggests that 'The more dramatic the misrepresentation, the greater the temptation to say the target is absent; but it's plainly open in any such case to say either.' (2004, p. 32) Perhaps, then, it is a mistake to think that we can establish whether self-misrepresentations have a real referent, and perhaps it is a mistake to think there is even any fact of the matter about their reference. One could frame matters in terms of a real object that differs radically from how we represent it, or one could equally well frame matters in terms of those representations having a merely intentional object. This would mean that the case for the illusion model fails, and that we are only justified in hovering over the disjunctive conclusion that the self is *either* merely intentional *or* real and misrepresented.

5.2.2 Reply

The first thing to note is that this objection only applies to my claim, developed in Sect. 4, that the illusion model is preferable to the hallucination model. The claim in Sect. 3 that the arguments for VST fail to rule out the illusion model still stand, for if it is indeterminate whether self-representations have an object then one cannot conclude that the self is a merely intentional entity. A less concessive response to the objection is that self-representations do not present the kind of indeterminate case proposed. It is one thing to countenance the possibility of such cases, but quite another to accept that the self presents such a case. I hope to have shown that the illusion model has clear theoretical advantages over the hallucination model. As such, we are not in a situation where theoretical considerations allow us to go either way.

5.3 The Indeterminacy of Selfhood

5.3.1 Objection

The previous objection was that it might be indeterminate whether our self-representations have a referent. An alternative objection is that it might be indeterminate whether some entity qualifies as the self. We have already seen in this paper that our conception of self-hood is nebulous, and that there are no clear necessary and sufficient conditions of self-hood. The choice between the hallucination model and the illusion model comes down to the choice between regarding *B* as the self, despite the fact it lacks many of the properties we attribute to the self, or instead regarding *B* as the generator of self-representations that refer to a merely intentional entity. But if the concept of the self is indeterminate then there may be no fact of the matter about whether *B* qualifies as the self, and so no fact of the matter as to which model of the self is true.

5.3.2 Reply

We ought to acknowledge that our conception of the self is fuzzy, and that there may be cases in which it is indeterminate whether the concept applies to some entity. But fuzzy concepts can still admit of clear cases: acknowledging the fuzziness of baldness does mean doubting that some people are clearly bald. Fuzzy concepts can also admit of cases where we can say with sufficient confidence that some concept applies to that entity whilst acknowledging the possibility of error: we can be justified in claiming that a person is bald whilst acknowledging the possibility that they have enough hair to qualify as non-bald. I suggest that *B*'s qualification as a self falls into this second category. The fact that *B* is the bearer of our mental representations and, moreover, that it is the bearer of our purportedly *de se* representations, gives us good reason to regard *B* as the self. Like the case of the not-quite-hairless bald person, we should temper our conclusion with the admission that we are in uncertain territory, but this is quite consistent with having a justified conclusion that *B* is indeed the self.

6 Conclusion

A wealth of considerations indicate that the self is not as it appears to be. The dominant interpretation of these considerations is that there is no such thing as the self—that it is a merely intentional entity. I hope to have shown that the dominance of this interpretation is unwarranted. First, we do not have reason to prefer the hallucination model of self-misrepresentation over the illusion model. Second, we have defeasible but considerable reasons to prefer the illusion model over the hallucination model. Specifically, our self-representations purport to refer to the bearer of those representations, so the burden of proof is on the hallucination model to give us reason to doubt this appearance and regard the self as an intentional entity distinct from the bearer of our self-representations. Choosing between the illusion

model and the hallucination model requires us to draw on some subtle theoretical points, but the choice could not be more important: it is the choice between whether or not we exist.

References

Bayne, T. (2010). *The unity of consciousness*. Oxford: Oxford University Press.

Blackmore, S. (2012). *The self illusion*. Retrieved: January 27, 2017 from the science and non-duality website: https://www.scienceandnonduality.com/videos/the-self-illusion-susan-blackmore/.

Colacurcio, R. (2012). *The virtual self: Beyond the gap in Buddhist philosophy*. Bloomington: Xlibris Corporation.

Dainton, B. (2012). Selfhood and the flow of experience. *Grazer Philosophische Studien, 84,* 173–211.

Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole, & D. L. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (pp. 103–115). New Jersey: Lawrence Erlbaum.

Dickens, C. (1843). *A Christmas carol in prose, being a ghost story of Christmas*. London: Chapman & Hall.

Farrell, J., & McClelland, T. (2017). Editorial: Consciousness and inner awareness. *Review of Philosophy and Psychology*, *8*(1), 1–22.

Gassendi, P. (1658). Opera omnia (Vol. III). Lyon: Laurent Anisson and Jean Baptiste Devenet.

Geddes, A. (2013). Think twice, it's all right: Animalism, disunity and the self. *Proceedings of the Aristotelian Society, 113,* 371–380.

Hood, B. (2012). *The self illusion: Why there is no 'you' inside your head*. London: Constable & Robinson.

Johnston, M. (2004). The obscure object of hallucination. *Philosophical Studies, 120*(1–3), 113–183.

Johnston, M. (2012). *Surviving death*. Princeton: Princeton University Press.

Le Morvan, P. (2004). Arguments against direct realism and how to counter them. *American Philosophical Quarterly, 41*(3), 221–234.

Lewis, D. (1979). Attitudes de dicto and de se. *Philosophical Review, 88*(4), 513–543.

Lichtenberg, G. (1971). *Schriften und Briefe,* (Vol. ii). Munich: Carl Hanser Verlag.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. Cambridge MA: MIT Press.

Nietzsche, F. (1886). *Beyond good and evil*. New York: Vintage.

Peacocke, C. (2014). *The mirror of the world: Subjects, consciousness and self-consciousness*. Oxford: OUP.

Rosenthal, D. (2004). Varieties of higher-order theory. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 17–44). Philadelphia: John Benjamins Publishing Company.

Van Gulick, R. (2014). Getting it all together—Phenomenal unity and the self. *Analysis, 74*(3), 491–498.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience* (2nd ed.). Boston, MA: MIT Press.

Williams, B. (1978). *Descartes, the project of pure enquiry*. London: Penguin.