Pete Mandik

Beware of the Unicorn

Consciousness as Being Represented and Other Things that Don't Exist

Abstract: *Higher-Order Representational theories of consciousness — HORs — primarily seek to explain a mental state's being conscious in terms of the mental state's being represented by another mental state. First-Order Representational theories of consciousness — FORs — primarily seek to explain a property's being phenomenal in terms of the property being represented in experience. Despite differences in both explanans and explananda, HORs and FORs share a reliance on there being such a property as being represented. In this paper I develop an argument — the Unicorn Argument — against both HORs and FORs. The core of the Unicorn is that since there are mental representations of things that do not exist, there cannot be any such property as being represented, and thus no such property with which to identify either being conscious or being phenomenal.*

Introduction

A surprisingly wide variety of theories of consciousness agree on the following two points. The first is that transitive consciousness — being conscious *of* something — should have explanatory pride of place in a philosophical theory of consciousness. That is, the concept of *being conscious of something* will play a key role in explaining *other* features of consciousness — features like what differentiates conscious mental states from unconscious mental states and what makes some properties *phenomenal* properties, properties in virtue of which there is *something it's like* to instantiate them. Call this first

Correspondence:
MandikP@wpunj.edu

idea upon which a wide variety of theories agree the Centrality of Transitive Consciousness.

The second idea — call it the Representational Reduction of Transitive Consciousness — is that one is conscious of something if and only if one has a certain kind of mental representation of it. There is divergence of opinion about what details appropriately specify which kind of representation is the right kind, but there is agreement among the philosophers I am concerned with that representing something is a requirement on being conscious of it. Thus the essential core of agreement in the Representational *Reduction of* Transitive Consciousness is the Representational *Requirement on* Transitive Consciousness. The main task of this paper is to raise problems for theories that agree on the Centrality of Transitive Consciousness as well as the Representational Reduction of Transitive Consciousness. More specifically, the problems that I intend to raise are based on intentional inexistence. Central, then, is the fact that we are capable of consciously representing things that do not exist.

Two very popular approaches to consciousness that involve dual commitments to the Centrality of Transitive Consciousness and the Representational Reduction of Transitive Consciousness are Higher-Order Representational theories of consciousness and First-Order Representational theories of consciousness.

Higher-Order Representational theories of consciousness — HORs — primarily seek to explain a mental state's being conscious in terms of the mental state's being represented by another mental state.[1] First-Order Representational theories of consciousness — FORs — primarily seek to explain a property's being phenomenal in terms of the property being represented in experience.[2] Despite differences in both explanans and explananda[3], HORs and FORs share a reliance on there being such a property as being represented. In this paper I develop an argument — the Unicorn Argument — against both HORs and FORs. The core of the Unicorn is that since there are mental representations of things that do not exist, there cannot be any such

[1] See, for example, Lycan (1996) and Rosenthal (2005).

[2] See, for example Dretske (1995) and Tye, (1995; 2000).

[3] The divergence in explananda involves difference only in what are taken to be the *primary* explananda. While HORs treat state consciousness as primary, they still have something to say about phenomenal properties (namely that they are properties mental states are represented by higher-order mental states as having). Likewise, while FORs treat phenomenal properties as primary they still have something to say about state consciousness (namely that they are states *with* which, but not *of* which, one is conscious).

property as being represented,[4] and thus no such property with which to identify either being conscious or being phenomenal.

While I think many varieties of HORs and FORs are vulnerable to the Unicorn, in this paper I target just a few exemplars: David Rosenthal's Higher-Order Thought theory (HOT) and the FORs developed by Fred Dretske and Michael Tye. In sections 1 and 2 I spell out the targeted HORs and FORs, respectively, emphasizing their main motivations and their reliance on the notion of being represented. In section 3 I spell out the Unicorn Argument and offer some brief defenses of its more controversial premises.

In sections 4 and 7, I examine and reject proposals that HORs and FORs may save themselves from the Unicorn by embracing the Direct Reference hypothesis (DR). According to DR, there exist representations such that they have representational content only if that which they represent exists. The gist of my complaint against wedding HORs and FORs to DR is that the most plausible representations for consciousness and phenomenality are the least plausible candidates for DR.

In sections 5 and 8, I examine and reject proposals that HORs and FORs may save themselves from the Unicorn by rejecting their reliance on the notion of *being represented* and embracing instead the notion of *representing*. The gist of my complaint against these proposals is that they are inconsistent with the primary motivations of HORs and FORs: Transitivity and Transparency, respectively. Transitivity and Transparency are further dealt with in sections 6 and 9, respectively. In a concluding section I speculate as to the role a notion of representing will have in future theories of consciousness.

1. HOR: HOT

Central to all HORs is acceptance of what Rosenthal calls the Transitivity Principle, and what I shall call simply, Transitivity.

[4] While the Unicorn as directed against HOR's bears superficial similarity to a line of objection against HOR's raised by Byrne (1997) and Neander (1998), it is in fact a distinct line of thought. First and foremost, the Unicorn's reliance upon the denial of such a property as being represented distinguishes it from the Byrne-Neander objection. Further, the Unicorn is stronger and generalizes beyond targeting standard HOR's and, as will be detailed in a later section, targets self-representational theories of consciousness advocated by, e.g. Kriegel (2003) and Van Gulick (2004). While self-representational theories are arguably immune to Byrne-Neander (and have been advertised as such), they are not immune to the Unicorn.

> (TRANSITIVITY): A state is conscious only if one is conscious of this state.[5]

Transitivity is supposed to be an independently plausible principle concerning which mental states are the conscious ones.

One of the main lines of reasoning in favor of HORs is that such theories constitute proposals for how Transitivity is implemented. HOR theorists regard representation as a way in which someone can be conscious of something. Being conscious of something is (or is a kind of) representing something. Thus Transitivity's requirement on conscious states that their possessors be conscious of them is implemented by HOR's proposal that conscious states are ones that are appropriately represented.

According to Rosenthal's HOT version of HOR, one is in a conscious state by being conscious of it, and one is conscious of it by thinking about it. In other words, according to HOT, a state's being conscious consists (in part) in its being thought about. I include the parenthetical 'in part' because there is slightly more to a state's being conscious on Rosenthal's theory than simply being the target of a thought. But I am unconcerned with the something more here. I question whether thinking about a state has *anything* to do with its being conscious. For simplicity of discussion, then, I shall simplify the core claim of the HOT theory as the view that a state's being conscious consists in its being thought about.

The key ideas of HOT that make it vulnerable to the Unicorn include the following. (1) Some mental states have the property of being conscious. (2) According to Transitivity, a mental state comes to have the property of being conscious in virtue of one coming to be conscious *of* that state. (3) One comes to be conscious of a state in an appropriate way by having an appropriate thought about the state.

To see Rosenthal's commitment to (1)–(3) consider the following quoted material. We see a commitment to (1) and (2) expressed in the following:

> Intuitively, it's a distinguishing mark of conscious states that whenever a mental state is conscious, we are in some way conscious *of* that state. To avoid confusion, I'll refer to our being conscious *of* something, whether a mental state or anything else, as *transitive consciousness*. And I'll call the property mental states have of being conscious *state conscious* (Rosenthal 2005, p.235).

And we see a commitment to (3) (and (2) as well) here:

[5] See, e.g., Rosenthal (2005)

> Conscious states are simply mental states we are conscious of being in. And, in general, our being conscious of something is just a matter of our having a thought of some sort about it. Accordingly, it is natural to identify a mental state's being conscious with one's having a roughly contemporaneous thought that one is in that mental state (Rosenthal 2005, p. 26).

2. FORs

While HORs are most directly concerned with explaining state consciousness, FORs are concerned with state consciousness only indirectly and they account for it as a consequence of their account of what they are most directly concerned with: qualia AKA phenomenal properties. Tye and Dretske embrace the wide-spread view that phenomenal properties are those properties in virtue of which there is something it is like to have conscious states. Central to FORs is their further embrace of the transparency thesis or, as I shall call simply, Transparency.

> (TRANSPARENCY): When one has a conscious experience all one is conscious of is what the experience is an experience of.

Like Rosenthal, and in keeping with the Representational Reduction of Transitive Consciousness, Tye and Dretske interpret 'conscious of' as indicative of representation: being conscious of something involves mentally representing something. Thus, according to FOR, the properties determinative of what it is like to be in an experiential state are the properties represented by the state. When experiences are veridical, the properties determinative of what it's like *just are* the properties of the objects as they are correctly perceptually represented (Tye, 2000, pp. 46–7, 51; Dretske, 1995, pp. 73, 83–4). So, for example, as Dretske puts it:

> [Q]ualia are supposed to be the way things seem or appear in the sense modality in question. So, for example, if a tomato looks red and round to S, then redness and roundness are the qualia of S's visual experience of the tomato. If this is so, then … if things ever *are* the way they seem, it follows that qualia, the properties that define what it is like to have that experience, are exactly the properties the object being perceived *has* when the perception is veridical (Dretske 1995, pp 83–4).

Thus are qualia a certain kind of 'represented properties,' that is, qualia are defined as 'phenomenal properties — those properties that…an object is sensually represented…as having' and as properties *not* of the experience itself (Dretske, 1995, p. 73).

Regarding this latter point, that phenomenal properties are *not* properties of experiences, Tye writes:

> Visual phenomenal qualities or visual qualia are supposedly qualities of which the subjects of visual experiences are directly aware via introspection. Tradition has it that these qualities are qualities of the experiences. Tradition is wrong. There are no such qualities *of experiences* (Tye, 2000, p. 49).

What FORs are theories of, then, is the second-order property of being phenomenal. What distinguishes phenomenal properties from non-phenomenal properties is that only phenomenal properties are represented in a certain way. A ripe tomato has lots of properties, but when one of them gets represented in a certain way, it goes from being a mere property to being a phenomenal property. When I correctly represent in experience the redness of a red tomato, the property determining what it is like to have this experience is a property *of* the tomato — the redness — and it (the redness) takes on the second-order property of being phenomenal by being represented in a certain way. More precisely, for FORs, being phenomenal *just is* the property of being represented in a certain way. What 'a certain way' means for FORs is of little consequence for present purposes since my concern is with whether being phenomenal is identical to *any* way of being represented and it is an aim of the upcoming Unicorn argument to show that it is not.

We can distill from the above remarks the following key ideas of FORs that make them vulnerable to the Unicorn: (1) Some properties have the property of being phenomenal (that is, the property of being determinative of *what it's like* to have a conscious experience). (2) It follows from Transparency that the only properties we are conscious of when we have a conscious experience are properties that the experience is an experience *of*, not properties of the experience *itself*. (3) One comes to be conscious of properties in ways relevant to *what it's like* by having an appropriate experiential representation of those properties.

3. The Unicorn

For the purposes of the Unicorn Argument, it will simplify things without being unfair to the Unicorn's targets to characterize HOR and FOR as follows: HOR is true if and only if the property of being a conscious state is the property of being a represented state. FOR is true if and only if the property of being phenomenal is the property of being a represented property.

The conclusions of the Unicorn are that HOR is false and that FOR is false. Such conclusions follow, of course, only from certain additional premises. And the additional premises in question are of differing levels of contentiousness. Relatively less contentious is the premise that being conscious is a property of some mental states and being phenomenal is a property of some properties. Such a premise is sufficiently tame that I will spend no effort defending it, conserving effort instead for the most contentious premise of the Unicorn: There is no such property as being represented.

Before turning to further remarks concerning this premise, it will be useful to lay out the structure of the Unicorn.

> **P1**. HOR is true if and only if the property of being a conscious state is the property of being a represented state.
>
> **P2**. FOR is true if and only if the property of being phenomenal is the property of being a represented property.
>
> **P3**. Being conscious is a property of some mental states and being phenomenal is a property of some properties.
>
> **P4**. There is no such property as being represented.
>
> **C1**. HOR is false (there being no property of being represented for the property of being conscious to be identical to).
>
> **C2**. FOR is false (there being no property of being represented for the property of being phenomenal to be identical to).

One way of succinctly conveying *why* one should accept P4 is that it is one of the consequences of the thesis that we may mentally represent the nonexistent. Such a line of thought is a close relative of a familiar view in the metaphysics of intentionality. Such a view arises as one of the ways of resolving the inconsistent triad that constitutes what we might call '*The* Problem of Intentionality':

1) Relations can obtain only between relata that exist.
2) There exist mental representations of nonexistent things.
3) Representation is a relation between that which does the representing and that which is represented.

One way of resolving the above triad is by accepting 1) and 2) as premises in an argument for the negation of 3). That such an approach is the most favorable approach to the Problem of Intentionality is a view I share with philosophers such as Tim Crane and Uriah Kriegel.[6] There is insufficient space permitted for a full defense of this approach, and I have little of originality to offer in developing a full

[6] See, in particular Crane (2001, especially pp.22–8) and Kriegel (2007, especially pp. 307–12; 2008).

defense. Further, a full defense is here unnecessary. This is because what matters for present purposes is not whether this approach to intentionality is defensible against all objections, but whether it is defensible against all objections *available to HOR and FOR*. It will nonetheless be useful to briefly address considerations in favor of 1) and 2). However, before addressing such considerations, I will discuss the relation between 3) and the key premise of the Unicorn.

The view that most directly comes out of the negation of 3) is that representation is not a relation between representer and represented and it remains to do the extra work of deriving that there is no such property as being represented. The extra work may be achieved relatively simply by noting the differential plausibility of the mutually exclusive and jointly exhaustive options of viewing the alleged property of being represented as intrinsic and viewing it as relational. I consider it obvious that the lion's share of the plausibility goes to the relational option. Sticking with the more plausible view that if being represented were a property then it would be relational (more specifically, being constituted by a relation between representer and represented) and combining it with the denial of 3) yields the crucial premise that there is no such property as being represented.[7]

Against the argument in the previous paragraph one might press the following argument by analogy. I here quote an anonymous reviewer for this journal who proposes such a line of objection: '*[R]unning* pretty clearly isn't a relation, and yet a distance or a race can have the property of *being run* — it has this property just in case there exists a runner who covers the distance or completes the race *by running*.' Now, I will grant to the objector that there is a sense in which running pretty clearly isn't a relation. However, I take this to be due to (a) there being a sense of 'run' wherein it is an intransitive verb and (b) the natural assumption that corresponding to this sense is a monadic property of running. However, there are also some senses of 'run' wherein it is a transitive verb, senses which the *Merriam-Webster Online Dictionary* gives as 'to pass over or traverse with speed' and 'to accomplish

[7] It is worth stressing that there are many philosophical positions that treat representation relationally in some sense of 'relational' that are not at all targeted by the current line of thought. For example, one may hold a kind of 'short-armed' conceptual role semantics of the kind held to be most plausible for accounting for the representational content of concepts like 'and' whereby a representation has its content *solely* in virtue of relations it bears to other representations. Holding that representation is relational in this conceptual-role sense is fully consistent with denying that representation is relational in the sense central to the third item in the inconsistent triad.

or perform by or as if by running'.[8] I see no reason to avoid the natural assumption that corresponding to these transitive senses of 'run' are relations. Neither do I see any reason for avoiding the natural assumption that the obtaining of such relations is required for something's *being run*.

I should point out that I am not so much convinced that I have the right view here of running as I am convinced that there's little to be gained in engaging in further debate about it. I do not suppose that the science, metaphysics, or logic of *running* is sufficiently advanced to really substantiate either side of a debate over whether instantiating the property of being run requires instantiation of a relation of running. As such, I do not expect that the science, metaphysics, or logic of *representation* has much to learn from further attention to running.

Relations can obtain only between relata that exist.[9]

I here adapt a point of Kreigel's that I'm particularly fond of: just as it makes little sense to say of a monadic property that it can be instantiated though no existing particular instantiates it, so does it make little sense to say of a two-place relation that it can be instantiated without the existence of two particulars instantiating it (2007, p. 311). Some philosophers might be tempted to regard as a counter-example to this claim that I can be the same height as Julius Caesar who no longer exists. Another potential counter-example might be that I can be the same height as Sherlock Holmes who never existed (and, perhaps, never will). I will postpone for now discussion of alleged relations to fictional characters and restrict discussion in the present paragraph to alleged relations to historical figures. One sort of response to the Caesar case is to point out how easily it admits of paraphrase in terms of relations between existents. One sort of paraphrase affirms that Caesar and I both exist, but tenselessly at different times (Quine, 1960, pp. 170–3). Another paraphrase is counter-factual: *if* Caesar still existed, he would be the same height as me (Kriegel, 2007, p. 12). Another sort of response is to point out how, for the purposes of this paper, what matters are the sorts of responses available to HOR and FOR. HOR and FOR are not theories of *memory*, which involves the representation of the way things *were*, but of *consciousness*, which (according to HOR and FOR, at least) involves the representation of the way things *are*. The

[8] run. (2008). In *Merriam-Webster Online Dictionary*. Retrieved October 21, 2008, from http://www.merriam-webster.com/dictionary/run

[9] For Crane's (2001) remarks on our relating only to existents, see pp. 24–5. For Kriegel's (2007) remarks on this, see pp. 311–2.

case of Caesar is relatively uninteresting, then, since it would not count as a case in which I am alleged to relate to an inexistent *current* state of affairs.

There exist mental representations of nonexistent things.[10]

Consider the following as a brief defense of 2). Suppose that, contrary to 2), we do *not* mentally represent nonexistent things. A natural formulation of this supposition would be 'There exist no mental representations of nonexistent things'. Assuming that this supposition can be grasped in thought and doing so involves having a mental representation the content of which is that *there exist no mental representations of nonexistent things* results in self-defeat, since doing so would involve assuming the existence of a mental representation of nonexistent things, namely, the allegedly nonexistent mental representations of nonexistents.

It is worth noting here some further examples of representations of nonexistents as well as strategies available to defenders of the current line of thought for describing such representations. A rich set of examples is the set of mental representations with contents equivalent to false existentially quantified claims. Included is Russell's (1905) famous analysis of 'The present king of France is bald' as having a logical structure expressible as 'There exists an x such that x is the king of France, for all y, if y is the king of France, y is identical to x, and x is bald'.

It might seem puzzling to some philosophers to treat Russellian analyses as representations of non-existents. This puzzlement perhaps arises for the following reason. If one were attracted to the view that representation was like reference and that both involved a relation to that which is represented/referred to, that is, if one were to embrace 3), then one might be tempted to describe Russell's achievement as showing how 'The present king of France…' need not be regarded as a representation of a non-existent.

However, philosophers who *deny* 3) may adopt a strategy for describing representations of non-existents that would permit regarding Russellian analyses as representations of non-existents. This general strategy may be conveyed by first noting how one may paraphrase seemingly relational attributions of representations in a non-relational manner. One may paraphrase the apparently relational 'John is thinking of unicorns' either using an adverbial construction as in 'John is thinking unicornly' or 'John is thinking unicorn-wise', or using an

[10] For Crane's (2001) remarks on our mentally representing inexistents, see pp. 23–4. For Kriegel's (2007) remarks on this, see pp. 310–1.

adjectival construction as in ‘John has unicorn thoughts’. (See Kreigel (2007; 2008) for elaborations and defenses of the adverbial option, though I will not here take sides on the relative merits of averbialism vs. adjectivalism). By adopting a strategy of non-relational paraphrases, one is thereby free to both adopt the strategy of Russellian analysis and describe the analysed thoughts either adverbially (‘John is thinking present-king-of-France-is-bald-ly’) or adjectivally (‘John has present-king-of-France-is-bald thoughts’).

Someone might object that we bear relations to unicorns on the grounds that there exist unicorns in non-actual possible worlds. On such a view, thinking about unicorns would relate one to unicorns after all, but the relations in question are inter-world, not intra-world. At this point we need to just change the example from something like unicorns to something like square circles. There can be thoughts about square circles insofar as we grasp thoughts like the thought that necessarily, there are no square circles. Thus can there be thoughts about square circles even though there are no possible worlds in which square circles exist. Even if there are non-actual possible worlds and relations between thinkers in one and horned horses in others, no such relations would obtain between thinkers here and square circles anywhere, since there are no worlds populated by square circles. The failure of thinking about to be a relation in the square circle case will allow us to see that thinking about fails to be a relation in all cases. ‘Thinking about’ has the same sense in ‘thinking about square circles’ and ‘thinking about unicorns’ and if thinking about is not a relation in the square circle case, then it is not a relation in the unicorn case either.[11]

[11] That ‘thinking about’ or ‘about’ has the same sense in these different kinds contexts (contexts differing with respect to existence and non-existence as well with respect to contingent non-existence and necessary non-existence) may be demonstrated by showing the failure of various linguistic tests of ambiguity. I am grateful to Chase Wrenn (2006; personal communication) for all of the following points and examples. (The following explicitly concern the question of contrast between existence and non-existence, but I presume similar remarks to extend to a question of contrast between necessary and contingent non-existence). A *zeugma test* for ambiguity reveals that ‘in’ and ‘for’ are ambiguous when the test is applied, respectively to,

(1) She arrived in a limosine and a very good mood.

(2) Shelby is a senator for Alabama and the invasion of Iraq.

Similarly ‘about’ admits of ambiguity in certain contexts such as in

(3) The talk was about 45 minutes and intentionality.

However, in the uses of ‘about’ that concern me in this paper, ‘about’ fails the zeugma test for ambiguity. If we first consider the following contextual information…

(4) Ponce knew that there was a spy in his party, but he did not know it was Orcutt. As he drifted to sleep, he kept coming back to one question: would the spy keep him from finding the Fountain of Youth?

One technical concern that arises in evaluating the kinds of arguments here provided in support of premise P4 of the Unicorn is the question of whether such inferences are *formally* valid. A problem that arises here is that it is difficult to motivate a formalization of the relevant notions without begging the key questions at hand.

Consider, for example, the question of how to formalize ‘We mentally represent non-existents.’ It is difficult to see how to formalize it without thereby begging the question regarding the truth-value of ‘Mentally representing is not a relation.’ If we formulate ‘we mentally represent non-existents’ as

$$(\exists x)(\exists y)(Px \;\&\; {\sim}Ey \;\&\; Rxy)$$

where ‘Px’ is ‘x is a person’, ‘Ex’ is ‘x exists’, and ‘Rxy’ is ‘x mentally represents y’, then, in addition to whatever problems are raised by introducing an existence predicate, we have also introduced the problem that the question has been begged against ‘mentally representing is not a relation’.

Consider, however, that on the other hand, if we formulate ‘we mentally represent non-existents’ as

$$(\exists x)(\exists y)[Px \;\&\; Rx \;\&\; {\sim}(\exists z)(Uz)]$$

where ‘Uz’ is ‘z is a unicorn’ and ‘Rx’ is a predicate we construct by presuming a language of thought and an apparatus of thought-quotation giving us ‘x is thinking ‘$(\exists z)(Uz)$’’, then, in addition to whatever problems are raised by the fact that we are quantifying into the opaque context of thought quotation, we have the problem that we have begged the question in *favor* of ‘mentally representing is not a relation’. It should be clear, then, that any formalizations of the relevant notions are going to have to wait until the relevant issues are settled *informally*. I return now to my own contributions toward such an end.

It is important to keep in mind that I am not interested in defending the nonexistence of the property of being represented against all comers — there are indeed many and they have explored many features of

…we see that there’s nothing wrong with the following.

(5) He could not stop thinking about Ortcutt and the Fountain of Youth.

Applying Quine’s (1960, p. 129) test for ambiguity reveals an ambiguity of ‘bank’ when…

(6) First national is a bank, not a bank.

…can be read as non-contradictory (by reading the first ‘bank’ as ‘financial institution’ and the second as ‘river bank’). But Quine’s test reveals no ambiguity of ‘about’ in the plainly contradictory

(7) Ponce was thinking about the Fountain of youth, not about the Fountain of Youth.

I presume that the kind of unambiguous applications of “about” to both existents and non-existents in (5) and (7) may be similarly unambiguously applied to both contingent and necessary non-existents.

this issue. I am interested instead in defending the view that there is no good way for the targets of the Unicorn to reject this premise of the Unicorn.

One might attempt to oppose the Unicorn's most contentious premise by stating a case in terms of a notion of truth in fiction. One might, for example, embrace the following pair of views:

(1) There literally is no such person as Sherlock Holmes and it is no more true of Holmes that he does coke than that he smokes pot.

(2) There is a sense of 'true' whereby it is more true of Holmes that he does coke than that he smokes pot.

One might hold that the 'truths' about Holmes in (2) hold in virtue of it being true in the literal sense, true in sense (1), that Sir Arthur Conan Doyle wrote stories about Holmes doing coke but not true in sense (1) that Doyle wrote stories about Holmes smoking pot. There is thus a sense, then, in which Holmes, in spite of not existing, instantiates properties.

Note, however, how little this will help HORs and FORs. Those theories need it to be true (in sense (1)) that there is such a property as being represented. HORs want to explain, among other things, what it means for a mental state that exists to be conscious. And their explanation will be one that includes, among other things, that it is true in sense (1), that the state in question is represented. *Mutatis Mutandis* for FORs and their explanations of being a phenomenal property in virtue of being a represented property.

However, when things are represented, it is infrequently true of them in sense (2) that they are represented. Consider Holmes again. While it is true (2) of Holmes that he does coke, since it is part of Doyle's story, it is not true (2) of Holmes that Doyle wrote a story about Holmes, since that is not part of the story. (I must confess to not having read many Sherlock Holmes stories, but I'm relatively confident in guessing that Doyle did not engage in the kind of self-referential meta-fiction that might constitute an exception to my claim). Whatever sense might be made of the instantiation of properties by non-existents in terms of truth in fiction, it is not going to be the kind needed to block the inference against the existence of such a property as being represented.

One further way of defending the view that there's no such property as being represented is by the following argument by analogy. The idiom 'kick the bucket' means 'die' and implies no relation to any bucket. One can kick the bucket in the idiomatic sense with out there literally existing a literal bucket. Kicking the bucket, in the idiomatic sense of the phrase, *never* entails a relation to a bucket and this is true even in cases in which one dies while literally kicking a bucket or even dies *because* of literally kicking a bucket. (One might literally kick a bucket while barefoot and have the misfortune of connecting with a sharp poison-coated burr on the bucket's rim). We can summarize this by saying that since, in the idiomatic sense of kicking the bucket, kicking the bucket is something you can do even though no bucket exists, kicking the bucket in the idiomatic sense is not something you do to a bucket even in situations in which there happens to be a bucket. Another way to summarize this would be to say that when one kicks the bucket in the idiomatic sense, there is no such thing as the bucket that is thereby kicked, where 'thereby' is used in a logical, not a causal, sense and the latter uses of 'bucket' and 'kicked' are their literal uses. And this is true even in situations in which one idiomatically kicks the bucket while also literally kicking a literal bucket. A final way that we can summarize this is by saying that there is no such thing as the property of being the (literal) bucket that is (logically) thereby (literally) kicked when one (idiomatically) kicks the bucket. By analogy, then, there is no such property as the property of being represented.

Is it possible for one to affirm the point of the kind of argument exemplified in the previous paragraph but concede that there is nonetheless a representation relation instantiated when, for instance, I am thinking a true existentially quantified thought? For example, when there exists such a thing as the only coffee cup in my hand right now, and I think, for example, *the only coffee cup in my hand right now is empty*, is there room for a concession that I am in such a case bearing a representation relation to that coffee cup? The kind of concession I want to discuss here is considered by Kriegel (2007, p. 312; 2008, p. 84) in the following terms: while Kreigel denies that representation ever *constitutively* involves bearing a relation to that-which-is-represented he concedes that (perhaps[12]) sometimes representation *contingently* involves bearing a relation to that-which-is-represented.[13]

I think that there are grounds for resisting such a concession. These grounds can be put in terms of my coffee cup and my coffee cup thought. The problem is not that there is no contingent relation that my

[12] Kriegel (2007; 2008) seems keen on being tentative about this: he suggests that it *may be* the case that a non-constitutive, contingent representing-relation is sometimes instantiated.

[13] I am grateful to Anthony Dardis (personal communication) and an anonymous reviewer for this journal for pressing similar concerns.

cup thought enters into. The problem is (a) that there are many contingent relations between the cup and the cup-thought (e.g., they are both members of the ordered pair <cup, cup-thought>, one bears the *has more mass than* relation to the other), (b) there are many contingent relations between my cup thought and things other than the cup, and (c) there is no basis for singling out one relation as a *representing* relation the cup-thought bears only to the cup that doesn't just devolve into affirming a *constitutive* representing relation that the cup-thought bears to the cup. Consider how one would go about deciding which of the many contingent relations that the cup-thought enters into is to be regarded as the instantiation of a *representing* relation between the cup-thought and the cup. It seems that the temptation would be overwhelming to say that the one that counts as a representing relation is the one that is instantiated in virtue of the cup-thought being a thought *of the cup*. But this seems to run afoul of the denial that representations bear relations to the represented *constitutively.*

Consider again the discussion of 'kicking the bucket'. If Jones idiomatically 'kicks the bucket' on the occasion of literally kicking a bucket, does it really make sense to say that there's a *contingent,* though not *constitutive,* relation of kicking thereby born between Jones and the bucket *in the idiomatic sense of 'kicking the bucket'*? Which, of the many relations entered into by Jones's state of dying, including the many relations between that state and the bucket, deserves to be singled out as the *idiomatic* kicking of the literal bucket? I do not think there's any sense to be made of this. Nor is sense to be made of analogous remarks about so-called non-constitutive representation relations.

It is worth noting that the above views (i.e., that there is no such property as being represented, that representation is not a relation between representer and represented) are fully consistent with the following combinatorial view of the representation of non-existents. According to this view, the representation of non-existents like unicorns involves either a combination of representations of existing things or a representation of a combination of existing things. Either way of construing what it is that is combined, it involves at some level the representation of things that do exist. So, in the case of unicorns, the actually existing things referred to in this combinatorial view will be horses and things with horns. Whatever the merits of this combinatorial view of the mental representation of non-existents, it is fully consistent with the above views. Suppose we formulate the combinatorial view as the view that one can think of non-existents, for example, unicorns, only if one bears some relation *E* to some existents, horses and horned things that, though ‘uncombined’ are not unicorns, would be unicorns if appropriately ‘combined’.[14]

However, this view would not entail that *E* is the so-called representing relation, *R*, the relation between representers and representeds. One way of conveying the key point is that *E*, in this example, is a relation to the *union* of the set of horses and horned things, but whatever relation there is to unicorns must be a relation to the *intersection* of the set of horses and horned things. Nothing essential hinges on stating this point set-theoretically. The essential point may instead be conveyed by saying that even if it were a requirement on representing unicorns that one bear relation *E* to the mereological fusion of at least one horse and at least one horned thing, *E* would not be the same relation as the representing relation, *R*, since a unicorn is not merely the mereological fusion of a horse and a horned thing and I can represent a mereological fusion of a horse and a horned thing without thereby representing a unicorn.

In brief, the combinatorial view says that having a thought entails the existence of something thereby related to. However, it remains to be shown that, contra the above views, when I have a so-called thought about something that does not exist, what I’m really doing is having a thought about only something that does exist. It remains untouched by the combinatorial view, then, that we represent the non-existent. My main point in bringing up the combinatorial view is that one can adopt a view whereby representation is grounded in things that really exist without thereby automatically needing to disagree with the key premise of the Unicorn.

As this section draws to a close, it is worth emphasizing that much more can be said both for and against the Unicorn’s most contentious premise than can be realistically canvassed in the present paper. Aside from the brief remarks already made, additional discussion relevant to the evaluation of the Unicorn will emerge in the upcoming sections concerning *direct reference*. One line of objection to the view that we represent the non-existent involves embracing the theory of direct reference and says that, for some representations at least, they have representational content at all only if that which they represent exists. I will examine this sort of idea at greater length in connection with HOR in §4 and in connection with FOR in §7.

[14] This view might qualify as a kind of empiricism if *E* is construed as some kind of experiential relation between a perceiver and objects of sensory perception.

4. HOR + DR

The direct reference hypothesis (DR), as I will construe it for present purposes, holds that there are certain mental representations such that (a) two or more of these representations are about the same object if and only if they have the same cognitive significance and (b) these representations have representational content only if that which they represent exists.[15] The most promising aspect of DR with respect to defeating the Unicorn is (b). Combining HOR with DR entails postulating that all of the higher order representations relevant to explaining consciousness have representational content only if the states they are representations of exist. If HOR could be combined with DR it would be immune to the Unicorn.

However, at least in the case of HOT, HOR cannot be plausibly combined with DR. This is due to troubles that arise in connection with part (a) of DR. When we examine the most plausible examples of attributions of consciousness-conferring higher-order thoughts, we find that they give rise to opaque contexts inconsistent with DR.

To see these points, consider an example. Suppose that Jones has some mental state that is a candidate for state consciousness. Suppose, then, that Jones sees that x is red. In order for the state of Jones seeing that x is red to be a conscious state, according to HOT, Jones must have a higher-order thought about that state. It is useful to consider what attributions of that thought would look like. We might attribute the HOTs by saying that

(1) Jones believes that he sees that x is red.

or

(2) Jones believes of himself that he sees that x is red.

[15] There are several reasons why labeling this view 'direct reference' might be slightly misleading, but not in any ways that would significantly outweigh the ease of exposition this mild simplification serves. I briefly call attention to these reasons in this note. There is, of course, the point that 'direct reference' is usually utilized to describe views about language, not mental representation. Additionally, the view here discussed might be more accurately characterized as a just one kind of direct reference theory, what might be called a mental-representational version of 'Pure Millianism', to be contrasted against not only various versions of mental-representational analogues Fregeanism, but also other versions of direct reference theory such as what Caplan (2007) calls 'Sense Millianism' and Thau (2002) calls 'Guise Millianism'. Adopting a simpler taxonomy, one might simply describe the view I'm here calling 'direct reference' as a view Chalmers (2004) and Thompson (2006) describe 'Russellian' in contrast against the 'Fregean'.

Either way, by the time we get to 'he sees that x is red' we are well into an opaque context. [16]

Suppose that seeing that *x* is red is identical to having neural activity pattern number 67 in area v4 of cerebral cortex. Consider that if we replace 'sees that *x* is red' in (1) and (2) with 'has activity 67 in area v4' then we wind up with sentences that may very well have the opposite truth values of (1) and (2). This is not because seeing that *x* is red is not identical to having activity 67 in area v4. This is because Jones may very well lack appropriate neurophilosophical sophistication to believe of himself that he has activity 67 in area v4. DR requires the intersubstitutability *salvae veritate* of co-referring terms for the alleged relata. If the defender of transparent HOTs were to insist on the possibility of the above substitutions as *salva veritate*, then the following problem arises. If the meaning of a term is purely referential, and HOTs determine what it is like, and 'sees that *x* is red' and 'has activity 67 in area v4' are co-referring, then Jones's perceptual experiences would seem to him to be the neural activity pattern 67 in area v4. I suppose, however, that while Paul Churchland's experiences may seem neural to Paul Churchland, Jones's experiences need not seem neural to Jones.[17]

Perhaps a different way of attempting to wed HOT and DR is by construing consciousness-conferring higher-order thoughts as referring demonstratively. Such a construal would entail that Jones's state of seeing *x* as red is conscious only if Jones has a Higher Order demonstrative thought expressible by 'this is a state of seeing that *x* is red' where the demonstrative 'this' refers, if at all, to a state Jones actually has. If the demonstrative 'this' fails to refer, then 'this is a state of seeing that *x* is red' fails to express a consciousness-conferring higher-order thought because it fails to express any thought.

One consequence of a direct reference theory of demonstrative thoughts is that any difference in reference of 'this' gives rise to a difference in thought. Two occasions of thoughts expressible by 'this is an umbrella' would be occasions of thoughts with different contents if the two occasions of the demonstrative 'this' referred to numerically

[16] Contra a suggestion by an anonymous reviewer, the problem of opacity still arises even if the HOT theorist abandons (1) and (2) for 'Jones believes, of the state of seeing that x is red, that he has it'. In this suggested way of attributing a higher-order thought to Jones, the word 'it' is embedded in an opaque context.

[17] For further discussion of Paul Churchland's claims that brain states may be introspectible as such by people with sufficient neuroscientific training, see AUTHOR. For Churchland's original claims, see Churchland (1979).

distinct umbrellas. We might summarize this point by saying that directly-referring demonstrative thoughts are object-involving.

The object-involvement of singular thought makes *thinking about* a relation. The having of a singular thought (if there were such things) involves a relation between a thinker and an object. And insofar as thought turned out to be a relation, then I would be happy to concede that being thought about is a property. If there were such things as singular thoughts, then I would concede that there is such a property as being thought about insofar as there would be such a property as being the object of a singular thought. However, the object-involving nature of singular thought is going to be a hindrance, not an asset to a theory of consciousness. One consequence of object involvement is that differences in objects referred to by embedded singular terms yield differences in thoughts expressed by the embedding sentences. Thus, numerically distinct thinkers pointing to numerically distinct umbrellas think qualitatively, not just numerically, distinct thoughts expressible by 'this is an umbrella'. And this is so regardless of how otherwise similar the thinkers, the umbrellas, and the respective environments of each thinker-umbrella pairing happen to be. If we were focusing instead on the thought 'Dogs are mammals', the numerically distinct thinkers mentioned above would have merely numerically distinct thoughts. The thoughts are, qua thoughts, qualitatively identical, for they have the same content, namely that dogs are mammals. (I am assuming of course, that particular thoughts are particular mental events). We are approaching the main problem for the singular thought proposal as an adequate account of consciousness.

The object-involvement of demonstrative thoughts does not fit well with the HOT theory. The main problem arises because, on Rosenthal's HOT theory, the contents of HOTs are supposed to be responsible for determining *what it's like* to have conscious states. Rosenthal states the relation between HOT and what it is like as follows:

> What it's like for one to be in a qualitative state is a matter of how one is conscious of that state. If I am conscious of myself as having a sensation with the mental quality red, that will be what it's like for me, and similarly for every other mental quality. And how we are conscious of our qualitative states is a matter of how our HOTs characterize those states. There being something it's like for me to be in a state with a particular mental quality is a matter of my having a HOT that characterizes a state I am in as having that mental quality (Rosenthal, 2005, p.186).

In other words, what it's like to be in a conscious state is one and the same as how one's state appears to one. Further, how the state appears to one is a matter of how the state is represented by a higher-order thought.[18]

The appearance-determining aspect of consciousness-conferring higher-order thoughts is the aspect that makes them so poorly modeled by demonstrative thoughts. Mere numerical differences can suffice to give rise to differences in demonstrative reference. However, mere numerical differences do not suffice to give rise to differences in appearance.

My physical doppelganger who lived on a physical doppelganger of the planet I live on, with a physically similar life history would, I take it, have conscious states such that what it is like for him to be in those states is like what it is like to be in mine. His object-involving thoughts, however, would differ from mine insofar as his ‘this’’s pick out a distinct umbrella from mine, his ‘here’’s distinct places, his ‘I’’s a distinct person. But just as his umbrella may very well appear just as my umbrella does, so will his lower-order mental states appear to him as mine do to me. Thus, in spite of diverging in the contents of our demonstrative higher-order thoughts, what it’s like to be me may very well be just like what it’s like to be my physical doppelganger.

I would like to briefly consider whether HOP (Higher Order Perception) theorists[19] can take advantage of a certain feature of perception to immunize their version of HOR against the Unicorn. Perception is frequently taken to be factive: one can only be truly said to perceive an elephant if one is in a certain kind of causal interaction with an actually existing elephant. Hallucinations of elephants may be perception-*like* or *quasi*-perceptual, but only actual elephants may be the targets of elephant perceptions.

What typically matters in debates between HOT and HOP is whether the higher-order representations that matter are, as the HOP-heads have it, analog and/or non-conceptual or, as the HOT-heads have it, more like beliefs, that is, assertoric propositional attitudes. What distinguishes HOP and HOT is irrelevant for the current point. Strip away these distinguishing features and we are left with nothing not addressed in the current paper: the suggestion that consciousness is conferred to states via a representation relation.

[18] See, e.g., Rosenthal (2002, p. 241).

[19] See, for example, Carruthers (2004).

5. Altered HOT and What It's Like

In previous sections, I have considered ways in which HOT advocates might try to argue against the Unicorn argument. They could attack the premises or they could make minor adjustments to HOT, such as wedding HOT and DR. In this section I consider possible major revisions to HOT.

One such revision is as follows. Perhaps HOT advocates can say that when one has a HOT one is thereby in a conscious state, regardless of whether there exists some target state to bear the property of being thought about. Such a move may result in a theory that is immune to the Unicorn. However, advocates of such a move need to ensure that the resultant theory still has a credible answer to the question of *why* certain states are to be regarded as conscious. Further, switching to such a theory may be ad hoc if it cannot be given an independent motivation.

One such independent motivation for this revision (which may also position the resultant theory to answer the relevant 'why' question) begins by embracing what I shall call 'the what it is like principle' or just WIL:

> (WIL): A conscious state is a state of a creature in virtue of which there is something it is like to be that creature.

WIL is very closely related to Nagel's view that 'an organism has conscious mental states if and only if there is something that it is like to *be* that organism' (Nagel, 1974, p. 436). Rosenthal thinks that HOTs are the states in virtue of which there is something it is like to be that creature. If Rosenthal were to accept WIL, it would turn out then that HOTs are the conscious states.

Another route to thinking that HOTs are the conscious states would be to say that whatever properties are instantiated in virtue of there being HOTs, they are properties *of* the HOTs, not properties of what the HOTs are about (because being thought about is not a property).

However we arrive upon the imagined revised HOT theory, we run into trouble. The main motivation for the HOT theory is that it provides an implementation of Transitivity. Transitivity says that one's mental state is conscious only if one is conscious of it. The 'of' of 'conscious of' is interpreted as the 'of' of intentionality — the 'of', for example, of 'thinking of'. If the HOT theorist is going to respond to the Unicorn argument by stating that the HOTs are themselves conscious, then we do not have an implementation of Transitivity. An implementation of Transitivity would have states of a subject being conscious in virtue of the subject being conscious of them. But HOTs are not about themselves, nor are they necessarily represented by anything else. Conscious unrepresented HOTs do not implement Transitivity.

It will not do for the HOT theorist to change their theory to have the consciousness of a state consist in its *representing* instead of its *being represented*. If a state's consciousness consists in its representing then it is not an implementation of Transitivity, but instead an implementation of something like:

(C): A state is conscious only if it makes one conscious of something.

But C is a long way from Transitivity. It more closely resembles remarks that FOR defenders such as Dretske have made *against* Transitivity. Dretske, in objecting to Transitivity, says that conscious states are states 'we are conscious *with*, not states we are conscious *of*' (Dretske, 1995, pp. 100–1). Principle C identifies a state's consciousness with its being a state we are conscious with.

Others argue that conscious mental states are conscious in virtue of being representations of, among other things, themselves, thus do they defend Same-Order Representational theories (SORs).[20] Principle C identifies a state's consciousness with its being a *consciousness of*. All that's added by the revised SOR is to add that what it must be a consciousness of is itself. However, given that being represented is not a property, it is not clear that what is added is anything at all. If *consciousness of* is still going to be analyzed as *representation of*, then even revised SOR is defeated by the Unicorn.

6. Living Without Transitivity?

The conclusion of the Unicorn argument is incompatible with HORs and HORs derive much of their plausibility from Transitivity. If the lesson of the Unicorn is something that we can live with, then perhaps we must learn to live without Transitivity or find a way of accepting it while rejecting HORs. However, accepting Transitivity while rejecting HORs seems unappealing. It is hard to see how Transitivity doesn't just lead to HORs. Resistance to abandonment of Transitivity may be due to the fact that Transitivity is intuitive and useful. However, I think that its intuitiveness can be explained away and its utility can be had by a substitute (see also AUTHOR).

[20] See, e.g. Kriegel (2003) and van Gulick (2004). See Weisberg (2008) for an excellent overview of SORs.

The intuitiveness of Transitivity can be explained away as due to the unnoticed triviality of the truth that every conscious state we are aware of is one we are aware of. If there are any conscious states that we are not aware of, such counterexamples to Transitivity would not be accessed from the first-person point of view, and any states so accessed would seem to prove the rule. The crucial question, however, is not whether every conscious state is one of which we are aware, but whether every conscious state is one that, *qua* conscious, *necessarily* is one of which we are aware. To see that this question may be answered in the negative, consider an analogous question with respect to trees. Even though we can (in fact, must) grant that every tree of which we are aware is thereby, trivially, a tree of which we are aware, we may still make sense of the question of whether every tree of which we are aware is one of which we necessarily are aware, or whether, instead, each tree that happens to intersect our awareness may have nonetheless done very well existing (*qua* tree) without our awareness. Many subscribe to a scientific word-view whereby the existence of trees predates the existence of awareness of trees. Why not similarly embrace a view of conscious states whereby they, contra Transitivity, predate our consciousness of them? Certainly, we shouldn't be prevented from doing so simply because of the *prima facie* intuitiveness of Transitivity, since the *prima facie* intuitiveness of Transitivity may be explained as due to the inaccessibility from the first-person point of view of conscious states of which we are unaware.

Transitivity is supposed to be the pre-theoretically compelling ground upon which HORs are based. However, if Transitivity did have as strong a pre-theoretical grip on us as HOR theorists would seem to think, then it should strike us as deeply odd to say of a mental state that *first* it was conscious and *then* we became conscious of it. If we are strongly in the grip of the Transitivity intuition, then we should feel compelled to describe one's conscious states as conscious only as or after we are conscious of them. However, consider a case in which a visual stimulus is popping in and out of consciousness, as in motion-induced blindness (Bonneh *et al.*, 2001) or monocular rivalry (Breese, 1899) experiments where visual information undergoes quite a bit of uptake by the cognitive system but all the while alternately goes into and out of consciousness. (It appears to the subject as though the stimulus is disappearing even though in reality the stimulus is constant). When the yellow dot or the horizontal bars pop back, how should one describe one's conscious percept? Is one compelled to say that the percept was conscious only as one became conscious of the percept? Or can one say without verging on nonsense that first the percept became conscious and only (a brief moment) afterward did one become conscious of the percept? If the latter option seems at all plausible, then Transitivity does not have a strong claim to pre-theoretic plausibility.

Perhaps the lovers of Transitivity would prefer not to defend it on grounds of its pre-theoretical plausibility but instead of its methodological utility. One nice thing about Transitivity is that it supplies or at least accords with a useful methodological criterion for deciding experimentally when one has a conscious state. In an experimental paradigm like monocular rivalry, we can ask the subjects when they are conscious of a particular visual stimulus and when they are not. If we are relying on the subject's report of his or her own conscious states, it seems as if we are relying on the subject's consciousness *of* his or her own conscious states. The states of a subject that the subject is unable to volunteer reports on are states that we regard both as unconscious states and states that the subject is not conscious of. Thus does Transitivity fit with an appealing experimental methodology.

However, we can reap the same benefits by what might be called 'deflationary transitivity'. We need not regard being conscious *of* a state as essential to its being *conscious*. We may instead regard being conscious of a state is a contingent reference fixer of 'conscious state'. We may introduce into discourse the class of conscious states as those states of which we are aware and leave open to investigation that what makes those states hang together as a kind is not our awareness of them, but something that they happen to share with states a person can have without necessarily being aware of them.

7. FOR + DR

In §4, I examined and rejected the proposal that maybe a kind of direct reference can save HOR theories such as HOT from the Unicorn. I want to do a similar thing here for FORs. The proposal of uniting FOR with DR raises special issues. One issue is that FORs concern representations of properties, not particulars. The second issue is that FORs concern representation in experience, not thought.

Recall that, in § 4, DR was described as holding that there are certain mental representations such that (a) two or more of these representations are about the same object if and only if they have the same cognitive significance and (b) these representations have representational content only if that which they represent exists. The question arises: what is the most straightforward way of adapting DR to fit with a theory of the representation of properties in experience? I think that (a) and (b) can serve as useful models. We can attempt to make suitable alterations, (a+) and (b+). The transformation of (a) into (a+) will obviously involve replacing 'object' with 'property'. Not so obvious is what to do with 'cognitive significance' although 'experiential significance' might suffice. Or, more to the point of a discussion of phenomenal consciousness, we may work with 'what it's like,' where sameness and difference in experiential significance may be regarded as sameness and difference in what it's like. Thus we have

> (a+): Two experiences represent the same property if and only if they are the same with respect to what it's like to have them.

Moving on to the modification of (b), the main problem to deal with is how to apply the exists/doesn't exist distinction to properties instead of objects. Two suggestions immediately arise. The first is to identify it with the instantiated/uninstantiated distinction. The second is to identify it with the possibly instantiated/necessarily uninstantiated distinction. I will focus on the second option, since I intend to present counter examples to FOR+DR and counter-examples to FOR+DR in terms of necessarily uninstantiated properties are *a fortiori* counter-examples to FOR+DR in terms of uninstantiated properties. Thus, part of what is entailed by combining direct reference with FOR is

> (b+): An experience represents a necessarily uninstantiated property if and only if there is nothing it is like to have the experience.

In what follows I will argue against the wedding of FOR and DR by arguing that there can be experiences for which there is something it is like but the represented property is necessarily uninstantiated.

We see (that is, visually represent) necessarily uninstantiated properties whenever we look at certain pieces of art by M.C. Escher. In many of Escher's artworks, we see what at first glance seem to be three-dimensional objects and their arrangements, but on further reflection couldn't possibly exist. For example, in Escher's 1960 lithograph, 'Ascending and Descending', we see (and thus visually represent) a finite set of stairs, each one of which is higher than some other.

Now, it is open for the FOR theorist to hold that what is paradoxical in viewing such a picture is restricted to what concepts one brings to bear on the experience and that the contents of the experiences themselves contain nothing contradictory because, for example, the contents of the experiences themselves concern only the representation of a distribution of shades of gray in the visual field. I don't think this response is particularly plausible, but I won't pursue this further here, for I think there are bigger and much more interesting problems for the FOR theorists, problems that arise from experiences with paradoxical contents not obviously attributable to any coinciding conceptual states.

Consider, for one such example, experiences of the motion aftereffect, or, more colloquially, the waterfall illusion. The effect occurs when one has been staring at a moving stimulus for a while, such as a waterfall, and then directs one's attention to a stationary object such as a rock wall. One will then undergo a paradoxical experience whereby one and the same object, the rock wall in this case, appears simultaneously to be moving and not moving.

The problem posed by the motion aftereffect is that it is a putative example in which the property experienced — the property of simultaneously moving and not moving — cannot be instantiated, for nothing in reality can be simultaneously moving and not moving. At this point, the FOR theorist may be tempted to re-describe the alleged experience in question as actually being *two* experiences, one of which is an experience of something as moving and the other of which is an experience of the very same thing as stationary. Such a move would block the attribution of representations in experience of necessarily uninstantiated properties. However, one might wonder what independent motivation can be provided for such a move so as to make it not so obviously *ad hoc*. Instead of dwelling further on the motion aftereffect, I would like to spend time on a class of examples even more powerful.

Due to peculiarities of the normal functioning of the visual system, we can experience coloured after-images. Readers are no doubt aware that after staring at a bright red spot and then directing their gaze at a white wall, they will experience a green afterimage. FORs provide a natural explanation of such after-images: though no green object need be present in the room, one undergoes so-called green afterimages in virtue of mentally representing in experience the instantiation of green in a certain region in space.

Under certain conditions, there can be induced in normal subjects afterimages with colours corresponding to no colour an object can have. Following Paul Churchland, let us call such colours 'chimerical colors' for they are 'color[s] that you will absolutely never encounter as an objective feature of a real physical object' (Churchland, 2005, p. 324).

The textbook case of an afterimage involves locating the afterimage on a white background by fixating one's gaze on a white wall or piece of paper. Chimerically coloured afterimages may be achieved when afterimages are located on non-white and non-gray backgrounds. For example, if one were to look at a pale-blue-green stimulus and then position the resultant orange afterimage on a maximally saturated orange background, the resultant afterimage will be coloured what Churchland calls 'hyperbolic orange' an orange which is 'more 'ostentatiously orange' than any (non-self-luminous) orange you have ever seen, or ever will see, as the objective color of a physical object' (Churchland, 2005, p. 328).

Locating afterimages on black backgrounds yields afterimages that *no* objects, self-luminous or not, could have. If one looks at a saturated yellow stimulus for 20 seconds and positions the blue afterimage on a black background, the resultant afterimage will still be blue but will be exactly as dark as black. This is especially interesting since, as Churchland points out, 'no *objective* hue can be as dark as that darkest possible black and yet fail to *be* black' (Churchland, 2005, p. 324). Even more interesting is what happens when one starts by looking at a saturated blue and positions a yellow afterimage on a black background. The resultant image is still yellow, but a yellow exactly as dark as black. This is especially interesting because we tend to think of yellow as a light hue. Ludwig Wittgenstein once asked '[W]hy is there no such thing as blackish yellow?' (1978, p. 106). The afterimages described by Churchland show that while there cannot be such a thing as blackish yellow, it may nonetheless be represented in experience. The representation of blackish yellow involves the representation of a necessarily uninstantitated colour, and as such, cannot be accommodated by any version of FOR wedded to DR.

Recall the sorts of objections the imagined FOR theorist raised against the Escher and waterfall illusion counterexamples to FOR+DR and note how ineffective such objections would be against the case of chimerically coloured afterimages. The objection against the Escher case was that the paradoxical contents were represented in conception, not experience. Whatever plausibility such an objection had in the case of viewing a picture of an ever-ascending staircase, it certainly has no plausibility in the case of coloured afterimages. The objection against the waterfall illusion was that perhaps what was happening was not a single experience of motion and its negation, but two distinct experiences, one of motion, and one of the lack thereof. Whatever plausibility such an objection had in the case of the waterfall illusion, it certainly has no plausibility in the case of coloured afterimages. It is quite clear that when one as an experience of a colour patch, even in the case of an afterimage, one is not undergoing three separate experiences, one each for the hue, the brightness, and the saturation of the colour in question. One is, instead, having a single experience, one which involves the representation of a single colour which, if instantiated, would also instantiate a particular hue, brightness, and saturation.

One possible FOR-friendly response would be to say that the necessarily uninstantiated properties described above are complexes of properties that are individually instantiable. Such a response would involve modifying FOR so that what it is like is solely determined by the atomic properties represented, not by their combination. But such a revision runs into a big problem, namely, that it makes binding irrelevant to what it's like. To see this point about binding, consider that there's a difference in what it is like to see (1) red squares and blue circles and (2) blue circles and red squares. However, the possible response under examination would make (1) and (2) subjectively indistinguishable, for the response under examination would make the sole determinants of what it's like the representation of redness, blueness, square-ness, and circularity.

8. Altered FOR and More of What It's Like

In § 5, I examined the proposal that HOR be modified in such a way as to make it immune to the Unicorn. I argued that whatever alterations might be possible would result in an account that no longer implements Transitivity, thus removing a central motivating consideration in favor of monitoring theories. The discussion in the current section will follow a parallel course where whatever modifications might save FOR from the Unicorn result in an account that no longer implements Transparency, thus removing the central motivating consideration for FOR.

One possible revision of FOR is as follows. Perhaps FOR theorists can say that when one has a certain kind of first-order mental representation, one thereby has a state that has phenomenal properties, regardless of whether there exists some object or properties that the state represents. Of course, in order for such a move to not simply be an *ad hoc* response to the Unicorn, it will need some sort of independent motivation. One plausible independent motivation for such a view would involve embracing a principle similar to WIL from §5:

> (WIL*): A phenomenal property is a property of a state of a creature in virtue of which (that is, in virtue of the property) there is something it is like to be that creature.

If there is something it is like in virtue of having a first-order mental representation, but there exists no object or property that is represented, then a natural suggestion is that the bearers of phenomenal properties are the first-order representations themselves, that is, the experiences themselves. Another route to this view is to say that whatever properties are instantiated in virtue of there being experiences are properties *of* the experiences.

Whatever route the FOR advocate takes to this proposed revision, serious questions arise as to whether we still have a theory that implements Transparency, since Transparency is read by FOR theorists such as Tye and Dretske as entailing that when one has a conscious experience, one is *not* conscious of properties of the experience itself. If these theorists were to change their minds and assert instead that when one has a conscious state one is conscious of properties of the experience itself, then this would constitute something very close, if not identical to, an embrace of Transitivity. I take it as pretty obvious that one cannot embrace Transitivity without abandoning Transparency. And if one embraces Transitivity, it is difficult to see how one can maintain allegiance to FOR.

9. Living without Transparency?

Much of the appeal and plausibility of FORs derives from Transparency. If the Unicorn argument defeats FORs one might conclude, given a prior commitment to Transparency, that this is more a problem with the Unicorn than with FORs. However, I think that Transparency is not as appealing a thesis as may appear at first glance.

Considerations against Transparency are perhaps best put in terms of a distinction between content and vehicle. Thus, when Transparency advocates assert that in introspection we may be conscious only of what an experience is an experience *of* and not any aspect of the experience *itself*, this may be translated into the vocabulary of 'content' and 'vehicle' by saying that we may be conscious only of the contents of experiences but not of the vehicles of the experiences. What are the vehicles of experiences? Different theorists give different answers. According to a Cartesian substance dualist, while my visual experience of a piece of wax may have as its content an extended unthinking thing, the vehicle of the experience, the experience itself, is a state of an unextended thinking thing. According to a psycho-neural type-identity theorist, the vehicle will be a state of an extended thinking thing, more specifically, a state of a brain. Elsewhere I have developed Churchlandish considerations against transparency and defended the thesis that brainstates may be introspectible as such (AUTHOR). I do not wish to recount those arguments here, but focus instead on the following points.

One big problem with Transparency is the assumption or assertion that content is not a property of the experience. Transparency intuitions are typically spelled out in ways that beg important questions about how to characterize the notions of truth and content as they apply to experience. The typical focus is on so-called veridical experiences. Suppose I have an experience as of seeing a red apple. Suppose further that the experience is veridical. The questions arise of what makes the experience an experience as of a red apple (the content questions) and what makes the experience veridical (the truth question). On one theory of content and truth, the content is a possible state of affairs, a possible co-instantiation of the properties of being an apple and being red, and what makes the experience true is if that possible state of affairs is actual. But there are various accounts of content and truth, not all of which yield the verdict that in accessing the content of an experience we are not accessing a property of the experience itself. For instance, if one were to embrace a role-theory of content and a coherence theory of truth, then one could further embrace the view that in introspectively accessing the content of a veridical experience, one is thereby accessing properties of the experience itself (albeit, relational properties of the experience).

In summary, if we are sometimes able to introspect brainstates, we are sometimes able to access vehicles. And as long as certain accounts of content and truth are live options, accessing the contents of experiences may be accessing properties of the experiences.

Why did Transparency seem plausible in the first place? Plausibly, we acquire concepts of objects in the external world long before we acquire concepts of experiences. Further, when we acquire concepts of experiences, prototypical examples are experiences *of* objects (as opposed to experiences of nothing at all). Further, the main *value* of experiences, our interest in them, concerns what they are experiences of. Unless I am perverse, upon having an experience of a bull charging at me, my first thought is to get out of the way of the bull, not reach for an anesthetic that will make the ensuing encounter with the bull fully un-experienced. While the above may be true, and may account for the plausibility of Transparency, none of it entails the literal truth of Transparency. None of it entails that it is impossible to introspect properties of experiences themselves.

In § 6, I noted how whatever usefulness was promised by Transitivity could be gained by what I called 'deflationary transitivity'. I would like to make a similar suggestion regarding Transparency. When we pick out a class of experiences as veridical visual experiences of red apples, being veridical visual experiences of red apples is a contingent reference fixer of a certain kind of brain state, a brain state that is neither essentially veridical nor essentially of red apples.

10. Conclusion

In this paper I have examined representational theories that have at their core the notion that being conscious or being phenomenal is to be identified with being represented. I have argued that since it is possible to represent things that do not exist, being represented cannot be a property, and thus, cannot be the property that being conscious (or being phenomenal) is identical to. I have argued that these kinds of arguments point out crucial flaws in the HOT version of HOR as well as in FORs.

The question remains of how and whether the notion of representation will play a role in the final theory of consciousness. While I have argued that the notion of *being represented* will have very little work to do, it remains open that *representation* may play a useful role in other ways. One possibility that merits further investigation is that the key notion will not be the notion of *being represented*, but instead, the notion of *representing*. One question that will need to be addressed if this topic is to be further investigated is the question of what is to motivate the view that consciousness can be explained in terms of representing.

In the various views discussed in this paper, there were various appeals to three main principles claimed to capture what is intuitively involved in having a conscious mental state: Transitivity, Transparency, and WIL. Due to the sorts of problems raised by the Unicorn, the worth of the first two principles should be viewed with much suspicion (though, as I have argued, way may usefully rely on 'deflationary' versions of Transitivity and Transparency). It remains, then, to give an account of how the notion of representing can best illuminate the notion of being in a mental state such that there is something it is like to be in it.

Acknowledgements

For especially detailed and useful feedback on various incarnations of the Unicorn Argument, I am grateful to Chris Gauker, Uriah Kriegel, David Rosenthal, Josh Weisberg, Chase Wrenn, Tad Zawidzki, and two anonymous referees. I am also grateful for feedback from Mason

Cash, Tanasije Gjorski, Paul Gowder, Colin McGinn, Eric Schwitzgebel and Brandon Towl. I am grateful, too, for feedback from audience members of a presentation of this paper in the Invited Symposium on Intentionality and Phenomenal Representation at the 2007 meeting of the Southern Society for Philosophy and Psychology in Atlanta, Georgia.

References

Byrne, A. (1997), 'Some Like it HOT: Consciousness and higher-order thoughts', *Philosophical Studies*, **86**, pp. 103–29.

Caplan, B. (2007), 'On Sense and Direct Reference', In M. Davidson (Ed.), *On Sense and Direct Reference: Readings in the Philosophy of Language* (Boston, MA: McGraw-Hill), pp. 2–16.

Carruthers, P. (2004), 'HOP Over FOR, HOT Theory', In R. Gennaro (Ed.), *Higher-order theories of consciousness: an anthology* (Amsterdam/Philadelphia: John Benjamins Publishing Company), pp. 115–35.

Chalmers, D.J. (2004), 'The Representational Character of Experience', In B. Leiter (Ed.), *The Future for Philosophy* (Oxford: Oxford University Press), pp. 153–81.

Churchland, P.M. (1979), *Scientific Realism and the Plasticity of Mind* (Cambridge University Press).

Churchland, P.M. (2005), 'Chimerical Colors: Some Novel Predictions from Cognitive Neuroscience', In A. Brook & K. Akins (Ed.), *Cognition and the Brain: The Philosophy and Neuroscience Movement* (Cambridge: Cambridge University Press), pp. 309–35.

Crane, T. (2001), *Elements of Mind: An Introduction to the Philosophy of Mind* (Oxford University Press).

Dretske, F. (1995), *Naturalizing the Mind* (Cambridge, MA: MIT Press).

Kriegel, U. (2003), 'Consciousness as Intransitive Self-Consciousness: Two Views and an Argument', *Canadian Journal of Philosophy*, **33**, pp. 103–32.

Kriegel, U. (2007), 'Intentional Inexistence and Phenomenal Intentionality', *Philosophical Perspectives,* **21***,* pp. 307–40.

Kriegel, U. (2008), 'The Dispensability of (Merely) Intentional Objects', *Philosophical Studies*, **141**, pp. 79–95.

Lycan, W. (1996), *Consciousness and Experience* (Cambridge, MA: MIT Press).

Nagel, T. (1974), 'What is it like to be a bat?' *Philosophical Review,* **83**, pp. 435–50.

Neander, K. (1998), 'The Division of Phenomenal Labor: A Problem for Representational Theories of Consciousness', *Nous,* **32** (S12), pp. 411–34.

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

Rosenthal, D. (2005), *Consciousness and Mind* (Oxford: Clarendon Press).

Russell, B. (1905), 'On Denoting', *Mind*, **XIV** (4), pp. 479–93.

Thompson, B. (2006), 'Colour Constancy and Russellian Representationalism', *Australasian Journal of Philosophy,* **84** (1), pp. 75–94.

Tye, M. (1995), *Ten Problems of Consciousness* (Cambridge, MA: MIT Press).

Tye, M. (2000), *Consciousness, Color, and Content* (Cambridge, MA: MIT Press).

Van Gulick, R. (2004), 'Higher-order Global States (HOGS): An alternative higher-order model of consciousness', In R. Gennaro (Ed.), *Higher-order theories of consciousness: an anthology* (Amsterdam/Philadelphia: John Benjamins Publishing Company), pp. 67–92.

Weisberg, J. (2008), ‘Same Old, Same Old: The Same-Order Representational Theory of Consciousness and the Division of Phenomenal Labor’, *Synthese,* **160** (2), pp. 161–81.

Wittgenstein, L. (1978), *Some Remarks on Color* (Oxford: Blackwell).

Wrenn, C. (2006), ‘Zeugma and intentionality’, Retrieved October 21, 2008, from http://conditionalmaterial.blogspot.com/2006/04/zeugma-and-intentionality.html