The APA has requested that the committee examine its charge or mission, so we will be reflecting on how we might modify our official charge. As the pages of our newsletter reveal, the community we serve has interests in the philosophy of artificial intelligence and computational cognitive science, the philosophy of information, issues in the philosophy of computer assisted pedagogy, and various ethical issues pertaining to the development and uses of computers, the Internet, robotic technology, and much more. There is some concern that the varied content of our newsletter might not be adequately reflected in the committee's current charge. For this reason, we will be examining how we might update our charge (mission statement), which currently reads as follows:

> The committee collects and disseminates information on the use of computers in the profession, including their use in instruction, research, writing, and publication, and it makes recommendations for appropriate actions of the board or programs of the association.

I encourage everyone who has suggestions about the charge of the committee to send them to me: mguarini@uwindsor.ca. Whether you think the charge should stay the same or be modified, we would like to hear from you. We also solicit any comments people might have about the name of the committee. Much appreciated if the comments could be submitted no later than December 31, 2016.

I look forward to working with my colleagues on the committee—Colin Allen, William Barry, Gary Mar, Fritz J. McDonald, Susan Schneider, Dylan E. Wittkower, and Piotr Boltuc—to serve the community of scholars interested in bringing philosophical reflection to bear on the wide range of issues involving computing and information sciences and technologies.

MIND ROBOTICS

Functionalism, Revisionism, and Qualia

Ron Chrisley
UNIVERSITY OF SUSSEX

Aaron Sloman[1]
UNIVERSITY OF BIRMINGHAM

1. REVISIONISM ABOUT QUALIA

Eliminativists about qualia (e.g., Dennett; Frankish, forthcoming) make this claim:

> NE: Qualia do not exist.

(For those that consider that wording paradoxical, NE can be glossed as "The term 'qualia' does not refer to anything.")

Some eliminativist arguments for NE proceed by first arguing for NP:

> NP: There is nothing that has (or: Nothing could have) all the properties that qualia realists take to be essential to qualia.

NE is then thought to follow from NP so obviously that the step is rarely, if ever, explicitly mentioned or justified. Some of Daniel Dennett's arguments against the reality of qualia can be seen as taking this form. "Quining Qualia" (1988), for example, employs such a strategy, the properties operative in the NP step being intrinsicness, ineffability, privacy, and immediacy. (In what follows, we will be assuming that these properties, as elucidated by Dennett, are indeed what qualia realists take to be constitutive of qualia. Much of what we have to say does not depend on this assumption.)

Revisionists, on the other hand, accept many or all of the arguments against there being features of conscious experience that are intrinsic, ineffable, private, and immediate, but depart from the eliminativists by not denying that qualia exist—with the proviso that qualia may not be what many people, (other) qualia realists and eliminativists alike, think they are. That is, revisionists hold NP but deny (or at least remain agnostic about) NE. (For ease of exposition, we will initially assume revisionists are qualia realists, but will return to the agnostic option in section 2.3.3.) In particular, revisionists deny that the NE follows from the NP. (How can that be so? We say more about that in section 2.1.)

Another way of expressing the difference between qualia revisionism and qualia eliminativism is in terms of the distinction between illusion and hallucination. Standardly, illusion is "any perceptual situation in which *a physical object is actually perceived*, but in which that object perceptually appears other than it really is,"[2] while the *Stanford Encyclopedia of Philosophy* defines a hallucination to "an experience which seems exactly like a veridical perception of an ordinary object *but where there is no such object there to be perceived*."[3] Thus, Blackmore: "To say that consciousness is an illusion is not to say that it doesn't exist, but that it is not what it seems to be more like a mirage or a visual illusion." So a reasonable alternative name for revisionism would be "illusionism." However, despite this widely accepted distinction between illusion and hallucination, some use the term "illusion" to include cases where, they claim, there is no object being perceived. For example, Frankish proposes "illusionism" as a name for the position "which holds that phenomenal consciousness is an illusion and aims to explain why it *seems* to exist."[4] According to the standard distinction, "hallucinationism" might be a more accurate (although perhaps less catchy) name for the position Frankish is advocating.

Some more examples of qualia revisionists may be helpful. Many (but not all) of those who embrace the "Grand Illusion" view of consciousness[5] are revisionists about consciousness in general, and some may be revisionists about qualia in particular. A particularly clear-cut case of a revisionist about qualia is Derk Pereboom; cf. his "qualitative inaccuracy hypothesis": "[I]ntrospection represents phenomenal properties as having certain characteristic qualitative natures, and it may be that these properties actually lack such features."[6] Another clear qualia revisionist is Drew

McDermott, who has explicitly embraced[7] the revisionist account of qualia put forward in our earlier work,[8] and which is restated here in sections 2.1 and 2.2.2. On the other hand, Michael Graziano's attention schema theory is hard to categorize as revisionist or eliminativist. Although in describing his theory he says things such as "awareness exists only as a simulation," which would put him in the eliminativist/hallucinationist camp, he also distances himself from such a simple metaphysical position:

> The attention schema theory could be said to lie half-way between two common views. In his groundbreaking book in 1991, Dennett explored a cognitive approach to consciousness, suggesting that the concept of qualia, of the inner, private experiences, is incoherent and thus we cannot truly have them. Others, such as Searle, suggested that the inner, subjective state exists by definition and is immune to attempts to explain it away. *The present view lies somewhere in between*; or perhaps, in the present view, the distinction between Dennett and Searle becomes moot. In the attention schema theory, the brain contains a representation, a rich informational description. The thing depicted in such nuance is experienceness. Is it real? Is it not? Does it matter? If it is depicted then doesn't it have a type of simulated reality?[9]

One last terminological twist is that Frankish uses the term "weak illusionism" to refer to revisionism as defined above:

> [Illusionism] should be distinguished from a weaker view according to which some of the supposed features of phenomenal consciousness are illusory. Many conservative realists argue that phenomenal properties, though real, do not possess the problematic features sometimes ascribed to them, such as being ineffable, intrinsic, private, and infallibly known. Phenomenal feels, they argue, are physical properties which introspection misrepresents as ineffable, intrinsic, and so on. We might call this weak illusionism, in contrast to the strong form advocated here.[10]

Frankish's definition of illusionism is helpful in highlighting a responsibility that both revisionist and eliminativist (illusionist and hallucinationist) accounts of qualia incur: the duty of explaining why things seem other than they are. For revisionists, however, this responsibility takes a form different from the eliminativist duty Frankish mentions. Even if technically correct, it would be misleading to describe the responsibility for the revisionist as that of "explaining why qualia *seem* to exist," since the standard reading of that phrase presupposes, unlike the revisionist, that qualia don't exist. Given that we are initially assuming that revisionists are realist about qualia, it would be more usual to describe their corresponding responsibility as that of explaining how we have knowledge of the existence of qualia. Beyond this, however, the revisionist needs to explain why qualia seem to have the properties that they seem to have, despite not having them. Carruthers, another revisionist, is very clear on this point:

> [A] successful explanation of phenomenal consciousness . . . should
>
> 1) explain how phenomenally conscious states have a subjective dimension; how they have feel; why there is something which it is like to undergo them;
>
> 2) why the properties involved in phenomenal consciousness should seem to their subjects to be intrinsic and non-relationally individuated;
>
> 3) why the properties distinctive of phenomenal consciousness can seem to their subjects to be ineffable or indescribable;
>
> 4) why those properties can seem in some way private to their possessors; and
>
> 5) how it can seem to subjects that we have infallible (as opposed to merely privileged) knowledge of phenomenally conscious properties.

Note that the first constraint does not have the "explain why it seems that" form the others do. This is important, as it highlights a possible explanatory advantage of the revisionist strategy as compared to the eliminativist one. The advantage concerns dealing with the worry: "How can consciousness be a hallucination, since only a conscious subject can suffer from a hallucination?" This is not the place to give a full assessment of this worry and responses to it, but the basic point we wish to highlight here is that in some situations, the revisionist view has more room for maneuver in replying to objections than does the eliminativist view. For example, consider L:

> L: A subject has qualia iff there is something it is like to be that subject.

Perhaps some qualia eliminativists would reject L. (For example, it might be that the only sense they can attach to "there is something it is like to be X" is no different from the sense of "X is conscious," though more obscurely expressed, and yet they are not eliminativists about consciousness.) But suppose for the sake of argument that both a qualia revisionist and a qualia eliminativist agreed on L. Then it follows that the qualia eliminativist must deny that there is something it is like to be a subject. And this can indeed be hard to square with also believing that consciousness is a hallucination, since it seems that only someone for whom it is like something to be them can suffer from a hallucination. But for revisionists, things are not so problematic. Yes, only someone for whom it is like something to be them can suffer from an illusion. But since revisionists do not deny that there are qualia, they can accept L and still hold that it is like something to be a subject, and thus that subjects can be victims of illusions (and hallucinations), including the illusions that qualia are intrinsic, immediate, ineffable and private. So, at least in some cases, the revisionist (illusionist) does not run into self-defeating trouble with the claim that consciousness is an illusion in the way the eliminativist (hallucinationist) runs

into self-defeating trouble with the claim that consciousness is a hallucination.

Returning to Carruthers' explanatory desiderata: Eliminativists (hallucinationists) will have similar explanatory obligations, but given the existentially negative nature of their position, two changes would have to be made to Carruthers' criteria:

1) Constraint 1 would likely need to be converted into the same "explain why it seems that" format as constraints 2–5.

2) Eliminativist obligations are not well expressed in language that presupposes the existence of qualia and the properties of qualia. Instead, they are more easily stated in terms of explaining the subject's linguistic behavior.

Thus we would have as desiderata the requirements to explain why people say such things as:

1) "Phenomenally conscious states have a subjective dimension," "Phenomenally conscious states have feel," and "There is something which it is like to undergo phenomenally conscious states"

2) "Phenomenal consciousness is intrinsic and non-relationally individuated"

3) "The properties distinctive of phenomenal consciousness are ineffable or indescribable"

4) "The properties distinctive of phenomenal consciousness are private to their possessors"

5) "We have infallible (as opposed to merely privileged) knowledge of phenomenally conscious properties"

Which is, in essence, the heterophenomenological approach (Dennett). Revisionism can therefore be viewed as a kind of *ontologically conservative* heterophenomenology:[11] in explaining people's (especially philosophers') qualia talk, do not assume that qualia have the properties that people attribute to qualia in such talk (that's the heterophenomenological part), but *do* assume (or at least leave open the possibility; see section 2.3.3) that the features of experience that people (incorrectly) attribute those properties to, namely qualia, do exist (that's the ontologically conservative part).

By highlighting Carruthers' desiderata, we do not mean to suggest that they are the only constraints on a satisfactory theory of qualia. A naturalistic theory of qualia of the sort we aspire to should not merely attempt to specify what qualia are and why they seem to be the way they seem, but should also explain how instances could have been brought into existence by natural processes occurring on an initially lifeless planet and how many intermediate forms of consciousness (and qualia), and supporting mechanisms (physical and virtual machinery) were required both in the evolutionary history of current highly conscious and intelligent organisms, and in the individual developments between a newly fertilized egg and the adult crow, monkey, squirrel, elephant, or philosopher.

Although these "extra" constraints will not play a central role in this paper, we should clarify one thing before moving on. In taking on board these biological constraints, we do not thereby commit ourselves to the view that only biological organisms can be conscious, have qualia, etc. On the contrary, we believe that ideally, a theory of consciousness should explain how, in principle, artificial intelligence products, such as future household robots, could also have various forms of consciousness, possibly including visual and tactile qualia, for example, and whether this could be implemented in current digital technology or whether some other sort of implementation would be needed (e.g., based partly on chemical computation, which Turing suggested was true of brains[12]).

Finally, any revisionist account of anything, qualia included, has to deal with charges of changing the subject. In the case of qualia, opponents of revision (eliminativists and realists alike) might insist that "qualities of experience that are ineffable, immediate, intrinsic, and private" is just what we *mean* by "qualia." So whatever a qualia revisionist is talking about (defending, explaining, etc.), they are not talking about qualia. We will discuss how two different revisionist accounts of qualia attempt to repel these charges in 2.1 and 2.2.1. It is to these accounts that we now turn.

2. FUNCTIONALISM AND REVISIONISM

With the revisionist strategy in view, in what follows we would like to clarify it further by comparing two functionalist revisionist accounts of qualia: our own proposal, which can be called "Virtual Machine Functionalism" (or VMF),[13] and Gilbert Harman's account.[14]

2.1 THE VIRTUAL MACHINE FUNCTIONALISM ACCOUNT OF QUALIA

Technically, the VMF proposal isn't revisionist in the sense expounded in section 1 (the reasons why not will be made clear in section 2.2.3). But the VMF account *does* embrace the key (ontologically conservative) revisionist belief that NE does not follow from NP.

The VMF approach assumes that there are various working designs for information-processing architectures for more or less intelligent (or at least competent) systems (i.e., organisms, or, possibly, artificial systems), some of which allow the system to attend to and acquire information about some of the intermediate data-structures involved in processing sensory information, and to discover differences between changes that are produced by changes in the physical environment and changes that result from changes in the perceiver—e.g., alterations of viewpoint, looking through distorting lenses, screwing up eyes, tapping lower eyelid, or developing new introspective capabilities, e.g., as a result of attending art school, or engaging in systematic self-observation.

Not all such discoveries are available for all systems or for all intermediate information structures. Some sensory details may be constantly overwritten, and in some cases, although they are used for online control in sensory-motor control loops, it may be that no records of the intermediate states are made available for "higher level" cognitive processing, or preserved for later inspection. For example, some of the internal states and processes of feature-detectors used for high-speed control of actions may be inaccessible to scrutiny. This would imply that changes in such states cannot be detected. The same goes for many information processes involved in metabolic functions (in normal circumstances, though, some of them change during infections and the changed states become detectable, e.g., during an attack of flu).

Moreover, the VMF approach allows that there may be several intermediate levels of abstraction in sensory/perceptual or motor processing, some but not necessarily all of which may be accessible to internal self-monitoring. This is obvious in language understanding and production (e.g., acoustic, phonological, phonemic, morphemic, lexical, and various syntactic, semantic, and pragmatic levels of processing). Only expert linguists are (or can easily become) aware of all of them, though all normal language users use them all. It may be possible for some individuals to develop various new sub-skills if they have extendable/trainable portions of their information processing architectures. However, these abilities are not all there from birth, and how the required mechanisms (architectural layers) develop is mostly unknown.

The heart, then, of the VMF account of qualia is the proposal that qualia are properties of the virtual machine states or components of those states that give rise to qualia talk (or qualia thoughts). It may seem, to the subject whose currently running virtual machinery includes such states or sub-processes, or data-structures, that these properties are immediate, intrinsic, ineffable, and private, but (the VMF account proposes that) such a subject is incorrect, and the fact that these properties seem that way to the subject in which they are manifested can be explained in terms of their informational properties (for details, see 2.2.1 and 2.2.2). This is the sense in which the VMF account of qualia is a revisionist one.

A further attraction of the VMF account, which we can do no more than note here, is its potential to integrate its constitutive and revisionistic explanations of qualia with explanations of their phylogenetic and ontogenetic origins and dynamics, which we proposed as being further constraints on a naturalistic account of consciousness in section 1.

As also pointed out in section 1, any revisionist account of anything, qualia included, has to deal with charges of changing the subject. The proponent of the VMF account is free to reply that to make that charge against them would be to confuse meaning and reference. Obviously, one can use different concepts (meanings) to talk about (refer to) the same thing. The revisionist is proposing we use different concepts to talk about a previously talked about subject, and is changing the subject only if those concepts do not preserve reference. The VMF account can ensure sameness of reference by relying on a causal theory of reference: it is hypothesized that the word "qualia" refers to whatever virtual machine states, substates, and processes cause and regulate our use of that word. Those virtual machine components can also be referred to by using the terms and concepts of a sufficiently accurate and detailed architectural account of the subject in question.[15] In such a case, co-reference is preserved, and so revision without changing the subject is accomplished.

It should be stressed that this model of scientific progress (a causal theory grounding sameness of reference to a subject matter in the face of a shift from a less correct to a more correct conceptualization or theory of that subject matter) is hardly new.[16] It is a standard way to make sense of the notion that the ancients had an incorrect account of the same stuff that our account of gold is of, rather than having a correct account of something else (since they had different concepts than we have now). What is more likely to strike some as novel is the application of this idea to the case of qualia talk instead of, e.g., gold talk.

2.2 HARMAN'S ACCOUNT OF QUALIA AND COMPARISON WITH VMF

We turn now to a comparative discussion of Harman's account of qualia. There are some broad points of agreement between his account and the VMF account: both are functionalist and accept that qualia as standardly construed are problematic, either in themselves, or in their recalcitrance with respect to functionalist modes of explanation. And in both accounts it is the standard understanding of qualia which has to be given up, not functionalism or qualia themselves. That is, both accounts are revisionist in spirit. But there are some notable differences between them, some of which are revealed in their answers to three questions: "Are we aware of qualia?" "Are inverted qualia possible?" and even "Do qualia exist?" We now discuss the two accounts' answers to these questions, in turn.

2.2.1 ARE WE AWARE OF QUALIA?

A key part of Harman's account is brought to the fore in his response to a standard, qualia-based objection to functionalist accounts of consciousness:

> When you attend to a pain in your leg or to your experience of the redness of an apple, you are aware of an intrinsic quality of your experience, where an intrinsic quality is a quality something has in itself, apart from its relations to other things. This quality of experience cannot be captured in a functional definition, since such a definition is concerned entirely with relations, relations between mental states and perceptual input, relations among mental states, and relations between mental states and behavioral output.[17]

Harman's response centers on making a distinction between two kinds of features in play in experience:

- Features by virtue of which an experience has the content it has (call them *C-features*)

- Features that one is made aware of by virtue of having an experience (call them *A-features*)

Harman argues that these are typically conflated, but are in fact disjointed. An experience presents something (call it the object of the experience) as being some way, as having some feature, character, or quality. It is the object of experience and the features that experience represents that object as having that a subject is made aware of by virtue of having that experience. The experience does not, Harman argues, have that feature itself. Nor does it present itself as having that feature. So one is not, by virtue of having an experience, made aware of the features of that experience, or at least not the intrinsic features of that experience by virtue of which it has the content it has.

Harman then deems these C-features to be the intrinsic features or intrinsic character of experience, allowing him to conclude that we are not aware of the intrinsic character of our experiences. The reply to the qualia-based objection to functionalism then comes swiftly: "[S]ince you are not aware of the intrinsic character of your experience, the fact that functionalism abstracts from the intrinsic character of experience does not show it leaves out anything you are aware of."

However, the objection which Harman posed against himself did not invoke the experience of pain in one's leg or experiencing a red apple, but the more introspective cases of attending to those experiences. So while one may concede that Harman is right that in normal experience the intrinsic qualities of those experiences may be inaccessible, one might yet suspect that this is not true for the introspective case at hand. Nonetheless, Harman insists the introspective case is the same as the non-introspective case.[18] Thinking that they aren't, that introspection can somehow reveal the intrinsic features of experience in a manner similar to how one can inspect the features of a painting by virtue of which it has its content, is, he claims, to make a false analogy between experiences and paintings:

> Things are different with paintings. In the case of a painting Eloise can be aware of those features of the painting that are responsible for its being a painting of a unicorn. That is, she can turn her attention to the pattern of the paint on the canvas by virtue of which the painting represents a unicorn. But in the case of her visual experience of a tree, I want to say that she is not aware of, as it were, the mental paint by virtue of which her experience is an experience of seeing a tree. She is aware only of the intentional or relational features of her experience, not of its intrinsic nonintentional features. Some sense datum theorists will object that Eloise is indeed aware of the relevant mental paint when she is aware of an arrangement of color, because these sense datum theorists assert that the color she is aware of is inner and mental and not a property of external objects. But, this sense datum claim is counter to ordinary visual experience. When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features of her experience. Nor does she experience any features of anything as intrinsic features of her experience.

Harman concludes by underlining the generality of Eloise's case in a way that is meant to hit home:

> And that is true of you too. There is nothing special about Eloise's visual experience. When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, including relational features of the tree "from here."

We can now ask: in what sense, if any, is Harman's account revisionist? One indication that it is revisionist is that the account is susceptible to a particular criticism, a susceptibility that is characteristic of revisionist accounts. The criticism, first mentioned in section 1, is that it changes the subject. Naïve (that is, non-revisionist) qualia realists could object that, in the sense of "intrinsic character" they use to characterize qualia, it is *impossible* that one not be aware of the intrinsic character of one's experience—"intrinsic character" is precisely meant to pick out the A-features of experience. So even if Harman is right in claiming that the C-features and A-features can come apart, "intrinsic character," they might argue, should track the latter, not the former. For these naïve qualia realists, qualia may indeed be what give experiences the content they have. But it is more central to the notion of qualia that they are qualities of which the subject of an experience is aware. Harman is in effect claiming that naïve qualia realists are wrong that there is anything "mental" one becomes aware of when one introspects (NP), but denying that this means there are no qualia, since qualia are the (non-introspectable) intrinsic properties of experience.

Although a full discussion of this "transparent" view of qualia is not possible here, we can say that crucial phenomenological argument on which Harman relies (involving Eloise, above) is not persuasive, at least not to us. When we turn our attention to the intrinsic features of our visual experience, our attention is drawn, at least some times, to what we referred to as "features of the mode of perception."[19] For example, it is a feature of my mode of perception of the monitor in front of me now that there is more legible detail near my current point of fixation, and that this increased level of detail moves as my point of fixation changes. These are not features of the monitor, nor are they experienced as such, at least not when I turn my attention to my experience. More importantly, they are not experienced as features of the monitor itself, nor are they experienced at all in the absence of introspection. Another example is one's awareness of motion when one gently wiggles one's lower eyelid with a finger, while looking at the tree. Our sensorimotor systems are good at determining

whether changes to the sensorimotor manifold are due to changes in what is being perceived, or something to do with the changes in the perceptual apparatus/perceiver.[20] Is it so improbable that this distinction might make itself apparent in phenomenal consciousness?

This phenomenological counter-argument and alternative model of introspection is not meant to be a decisive refutation of Harman's view. Our phenomenological clash here is merely touching on a well-established debate between two views of introspection, the traditional "inner target" view, which can be traced back via Armstrong to Locke, and "transparency" views like Shoemaker's (and Harman's) that replace the idea that introspection is a kind of inner sense with the claim that it is rather a way of attending to the qualities of the perceived object (even if that object has to be an intentional object in the case of non-veridical, perception-like experiences). We do not presume to resolve this dispute here; rather, we wish to highlight this disagreement as a key difference between our functionalist account of qualia and Harman's. For those functionalists who do not wish to embrace the view that we are not aware of the intrinsic qualities of our experiences, there is an alternative.

Although functionalism and the "inner target" view of introspection are both well-known, traditional views in the philosophy of mind, they come together in the VMF account of qualia in a novel way. On the VMF account, when one introspects, one is having an experience[21] (N) the object of which is that (or another) experience (E), such that N represents E as having particular features, character, or qualities f. Further, it might be claimed, it is these features f of E that give E the content it has (i.e., that make it the case that E has an apple as its object and that E is presenting that object as being red). Harman may be right that a subject is not made aware of f by virtue of having E (which instead makes available an apple and redness). But, plausibly, one *is* made aware of f by virtue of having experience N, the introspection of E.

On the VMF account it is also the case that along with any experience E and features f of E that you are aware of by virtue of having introspective experience N, there will be many aspects of the information processing episode that you are (merely) potentially aware of (e.g., that you would become aware of if you reflected on other cases, or if something happened to draw your attention to differences between two experiences that involve changing relationships). For example, if you dimly experience a familiar face reflected in a window, you may fail to notice that part of the experience concerns the distance of the face. But you might come to notice that if the reflected face moved closer. For the VMF theorist, this merely points to the fact that the content of a vast amount of processing does not receive attention, but is capable of doing so, as distinct from other processing where the information used is beyond the reach of (normal) consicousness, e.g., low-level acoustic processing of speech sounds or visual processing of colors (which appears to be non-relational but is highly "relational" as shown by various illusions).

One might wonder how N could have E as its object. Given that N and E are distinct experiences, if a subject is having experience I, then she is ipso facto not having experience E, and thus, while the subject is having N, there is no experience E to serve as the object of N. At best, N can have as its object a memory or other representation of E that exists at the same time as N.

There is more than one way to respond to this worry. One response notes that the worry relies on the following assumption concerning the temporal relation between perception and the objects of perception (exteroception and interoception alike):

> T: For x to be the object of a perception at time t, x must exist at time t.

This assumption can be questioned. Of course non-existent objects cannot enter into relations, but that is not required here. All that is required is that a relation can hold at time t between an object that exists at time t and another object that exists at a time earlier than t. In fact, we find it natural to say that a subject is seeing a distant star (and not seeing a representation or memory of that star), even in the case where the star in question ceased to exist millions of years before the subject's birth.

Another line of response is to maintain that N and E *can* exist at the same time. For example, it might be that a subject can have more than one distinct experience simultaneously, or that experiences can have other experiences as proper parts. VMF is well poised to make sense of these proposals by way of identifying[22] experiences with components of virtual machine states and processes, given that it explicitly differs from standard functionalism in allowing for functional sub-states that can be tokened simultaneously, or nested.

However, if one still had doubts about these mereological possibilities for experiences, there is a third line of response that explicitly draws on features of the VMF version of revisionist functionalism in a different way. If qualia are identified with (or implemented in; see footnote 4) properties of virtual machine states, then it may very well be that one can only be having an experience with a given quale if one is in the corresponding VMF state. But it is possible to get information about, or "inspect" VMF states that are not tokened by inspecting the computational structures that are responsible for their deployment and implementation. So even if E must be tokened at t in order to perceive E at t, and even if having an introspective experience N precludes being in experiential state E at the same time, one can still make room for the "inner target" view of introspection by taking the relation between N and E to be intentional, but non-perceptual. By virtue of being in N one can be made aware of the features of E because N is causally related to the computational determinants of E.

By the computational determinants of a virtual machine state E we mean the currently tokened computational states and properties that, once a triggering condition for E's tokening is met, will jointly determine that it is E that is tokened, as opposed to some other virtual machine state E'. For example, my computer is not now running the Firefox

application. So the virtual machine state "running Firefox" is not now tokened by my computer. But the computational states and properties currently tokened by my computer include the hard disk memory states that store the code for Firefox. And it is these states (among others) that make it the case that, when I click on the Firefox icon, my computer enters into the "running Firefox" virtual machine state.[23]

It is worth noting that in general, some states might have some of their properties because their determinants (computational or otherwise) have the very same properties. Because of the relative abstractness of computational states, this is especially likely for virtual machine states and their determinants. This means that the VMF account of qualia can make sense of the introspection N of a not-currently-tokened experience E, even on a perceptual understanding of introspection. Even if there can be no perception of E itself, there can be perception of the features f of E via perception of the same features f of the determinants of E, together with the fact that, say, there is a law that ensures that if the determinants of E have f, then E will have f as well.

One advantage of the "inner target" character of the VMF model of introspection is that it does not require, unlike transparency accounts such as Harman's, an appeal to intentional objects to serve as the objects of experience, and therefore as the objects of introspection, in cases of imagination or hallucination. Recall that the transparent account understands introspection as becoming further acquainted with the qualities of the object of experience (e.g., a tree). When, as in imagination or hallucination, there is no physical object of experience, Harman's account requires that there be an intentional object of the experience, and it is the features of this intentional object, not of any experience, which one is aware of when one introspects in such situations. By contrast, if the VMF account has any explanatory connection with intentional objects, it is in the reverse direction; intentional objects are not used by the VMF account to explain anything, but rather the VMF account can be seen instead as explaining or naturalizing purported relations to such objects (or the temptation to speak as such) in terms of as relations to physically-realizable objects: virtual machine states.

To recap this section: Harman's revisionism is apparent in how he deals with a standard objection to functionalist accounts of qualia. Locating the problem in the notion that qualia are both the intrinsic features of experience *and* the objects of introspection, he dissolves the problem by asserting that qualia are the former and not the latter, implicitly asking us to revise our concept of qualia accordingly. He also attempts to explain why it seems to some (e.g., naïve qualia realists) that qualia are both. The VMF account of qualia, while revisionist with respect to other aspects of qualia, is neutral on this issue, being consistent with an "inner target" model in which the objects that introspection makes us aware of are indeed the intrinsic qualities of experience—properly understood.

2.2.2 ARE INVERTED QUALIA POSSIBLE?

Another well-known objection to functionalist accounts of qualia is based on the notion of spectrum inversion. Harman summarizes the problem:

> [I]t is conceivable that two people should have similarly functioning visual systems despite the fact that things that look red to one person look green to the other, things that look orange to the first person look blue to the second, and so forth (Lycan 1973, Shoemaker 1982). This sort of spectrum inversion in the way things look is possible but cannot be given a purely functional description, since by hypothesis there are no functional differences between the people in question. Since the way things look to a person is an aspect of that person's mental life, this means that an important aspect of a person's mental life cannot be explicated in purely functional terms.[24]

Harman introduces us to Alice and Fred, an inverted spectrum pair: "Things that look red to Alice look green to Fred, things that look blue to Alice look orange to Fred."[25] He then gives us a quick theory of perception in which perceptual representations, which have enough causal efficacy to serve as guides, play a central role:

> Perceptual processing results in a perceptual representation of that strawberry, including a representation of its color. [Alice] uses this representation as her guide to the environment, that is, as her belief about the strawberry, in particular, her belief about its color.[26]

Harman then offers a solution which has at its heart this:

> The hypothesis of the inverted spectrum objection is that the strawberry looks different in color to Alice and to Fred. Since everything is supposed to be functioning in them in the normal way, it follows that they must have different beliefs about the color of the strawberry. If they had the same beliefs while having perceptual representations that differed in content, then at least one of them would have a perceptual representation that was not functioning as his or her belief about the color of the strawberry, which is to say that it would not be functioning in what we are assuming is the normal way.[27]

Harman expresses this claim, that a difference of qualia must involve a difference in function, in another way:

> [T]here can be nothing one is aware of in having the one experience that one is not aware of in having the other, since the intentional content of an experience comprises everything one is aware of in having that experience.[28]

The critic of functionalism will no doubt find the forgoing unsatisfying. To assume that a difference in qualia amounts to or requires a difference of "perceptual representation" or "intentional content" in a sense that has any causal

relevance is to beg the question. In terms of the first passage just quoted, the critic of functionalism will insist that Harman needs to address the case in which the beliefs are the same *and* the perceptual representations are (functionally) the same, yet the qualia are different. Harman retorts that it is only someone who assumes that we are immediately and directly aware of the intrinsic features of experience who can plausibly imagine qualia floating free of perceptual representations and intentional content in this way. And to his lights he has already discredited that assumption (see section 2.2.1)—although we tried to sketch an alternative to his view.

The forgoing may or may not be a valid and/or novel criticism of Harman's position; whether it is any of those is subsidiary to the main purpose here, which is to compare and contrast Harman's account of qualia with the VMF account. Since we sketched a way that one might defend the "inner target" view of introspection, and since Harman diagnoses that view as being what enables a view of qualia that completely floats free of function, representation and intentional content, is the VMF account not in trouble? No—the "inner target" view of introspection might be necessary for naïve qualia realism, but it does not imply it, as we hopefully demonstrated in 2.2.1.

More important for a comparison of the VMF account and Harman on this issue is not the success or failure of his response to the inverted spectrum challenge, but that he accepts that it is a valid, well-posed challenge at all. Such acceptance is in stark contrast to the VMF account, which has the implication that, at least in the case of some qualia, it is incoherent to wonder if a quale in one individual may or may not be the very same quale as that in another individual. To assume at the outset that it makes sense, for any given quale, to compare it to a quale in another subject is to risk making a category mistake.

This might seem an odd claim to make. The VMF account identifies qualia with (properties of) virtual machine states, which are themselves public, objectively observable phenomena, so why can't their properties be compared or identified? Can't we ask (and answer) the question of whether two computers (say) are in the same virtual machine state? Things get notoriously problematic when comparing the functional states of non-functionally identical systems, but what about the functionally identical case? Surely when two systems are functionally identical, the question of whether or not they are in the same virtual machine state (and therefore have the same qualia) has a clear, positive answer?

Well, yes and no (a common revisionist response!). Yes, in that qualia are *actually* properties of objective, publically observable virtual machine states, they are comparable, can be re-instantiated, etc. They are not private or ineffable.[29] But this is not engaging with the critics of functionalism on their own terms, saying only this is unlikely to persuade a non-revisionist.[30]

To translate what the naïve qualia realist is concerned with into the VMF framework, one needs to consider not (just) architecture-based concepts, such as that of a virtual machine state, which assists the theorist in understanding the features of a cognitive architecture, including the properties of its experiential states. One needs also to consider what we call architecture-driven concepts, which are concepts the architect makes available to the subject that the architecture is an architecture of.[31] The architecture-driven concepts with which we are concerned here (the ones that will explain why qualia seem to be private and ineffable) are created within an architecture as part of the individual history of the architecture or machine. Now, suppose that agent A with a meta-management system uses a self-organizing process to develop architecture-driven concepts for categorizing (properties of) its own internal virtual machine states as sensed by internal monitors. If such a concept C is applied by A to one of its internal states (or one or more of its properties), then the only way C can have meaning for A is in relation to the set of concepts of which it is a member, which in turn derives only from the history of the self-organizing process in A. These concepts have what Campbell refers to as "causal indexicality."[32]

The implication of this is that A's qualia, *as experienced/represented by A*, are not the kind of thing which could be in a system other than A. If two agents A and B have each developed concepts in this way, then if A uses its concept C_a, to think the thought "I am having experience that is C_a," and B uses its concept C_b, to think the thought "I am having experience C_b," the two thoughts are *intrinsically private* and *ineffable*, even if A and B actually have exactly the same architecture and have had identical histories leading to the formation of structurally identical sets of concepts. A can wonder: "Does B have an experience described by a concept related to B as my concept C_a is related to me?" But A cannot wonder "Does B have experiences of type C_a?" for it makes no sense for the concept C_a to be applied outside the context for which it was developed, namely one in which A's internal sensors classify internal states. They cannot classify states of B. This privacy and ineffability of C_a it will likely make it seem to A that its experiences have properties (that is, the qualia represented by concept C_a) that are private and ineffable.

To reiterate, when different agents use architecture-driven concepts, that are produced by self-organizing classifiers, to classify *internal states of a virtual machine*, and are not even partly explicitly defined in relation to some underlying causes (e.g., external objects or a presumed architecture producing the sensed states), then there is nothing to give those concepts any user-independent content in the way that our color words have user-independent content because they refer to properties of physical objects in a common environment. Thus self-referential architecture-driven concepts used by different individuals are strictly non-comparable: not only can you not know whether your concepts are the same as mine, the question is *incoherent*. If we use the word "qualia" to refer to the (properties of) virtual machine states or entities to which these concepts are applied, then asking whether the qualia in two experiencers are the same would then be analogous to asking whether two spatial locations in different frames of reference are the same, when the frames are moving relative to each other. But it is hard to convince some people that this makes no sense, because the question is

grammatically well-formed. Sometimes real nonsense is not *obvious* nonsense.

So the naïve qualia realists win the battle: (some) *thoughts about* qualia are intrinsically private and ineffable. But they lose the war: qualia themselves are not intrinsically private and ineffable, only some ways of thinking of them are—the ways that are afforded by causally indexical, architecture-driven concepts of a particular sort.

Not everyone will be happy with our position here. For example, contrast our view with what Pete Mandik says in this passage criticizing Lycan's indexical response[33] to Jackson's Knowledge Argument:[34]

> One such problem with the indexical response is that it mistakenly makes numerical differences sufficient for subjective differences. To see why this is a bad thing, consider the following. Suppose that while Mary does not know what it is like to see red, Cheri, Mary's color-sighted colleague, does know what it is like to see red. Upon seeing red for the first time, not only does Mary learn what it is like to see red, she learns what it is like to be Cheri. If Mary and Cheri were physical and experiential doppelgangers (though numerically distinct individuals) they could each know what it is like to be the other person, regardless of whether their numerical non-identity entails divergence of the contents of their indexical thoughts.[35]

If what we are saying is correct, there is a sense in which Mary does not learn what it is like to be Cheri. On our view, even physical doppelgangers do not know, in this sense, what it is like to be their fellow doppelganger. Worse, in this sense, the notion of "experiential doppelgangers" is incoherent. Whether this point could be turned into a defence of the indexical response to the knowledge argument is a possibility we will have to consider on another occasion.

Harman acknowledges an explanatory gap "between some aspect of our conscious mental life and any imaginable objective physical explanation of that aspect."[36] But he rejects that this explanatory gap implies a metaphysical one, instead locating it in the difference between objective and subjective understanding. A functional account of what goes on when someone has an experience is an objective account and, Harman argues, cannot in itself provide understanding of what it is like to have that experience, which requires subjective understanding. In particular, one must be functionally similar enough to the subject one is trying to understand:

> Suppose we have a completely objective account of translation from the possible experiences of one creature to those of another, an account in terms of objective functional relations, for example. That can be used in order to discover what it is like for another creature to have a certain objectively described experience given the satisfaction of two analogous requirements. First, one must be able to identify one objectively described conceptual system as one's own. Second, one must have in that system something with the same or similar functional properties as the given experience. To understand what it is like for the other creature to have that experience is to understand which possible experience of one's own is its translation. If the latter condition is not satisfied, there will be no way for one to understand what it is like to have the experience in question. There will be no way to do it unless one is somehow able to expand one's own conceptual and experiential resources so that one will be able to have something corresponding to the other's experience.[37]

Recall that on the VMF account, there are some ways of thinking of (some) qualia that are, because of their history and causal indexicality, inherently private, non-shareable, and system specific. The implications of this are problematic for Harman's position as stated above. Let's assume that a subject A knows what it is like to be A, to have the experience A is now having. This knowledge, Harman would agree, consists in having the right conceptual resources to represent that knowledge. Whether B can know what it is like to experience what A is experiencing depends on what is to count as a proper "translation" of the concepts A is using. One could merely require the concepts to have similar functional profiles, which would yield Harman's position: B can understand subjectively what it is like to be A if B is functionally similar enough to A. But this will not impress the naïve qualia realist, who would maintain that sameness of functional role (even of concepts) is not enough to capture qualia (because we can imagine them coming apart). So to explain qualia in a sense that is at least continuous with the way the naïve qualia realist thinks of them requires a stronger notion of "translation." The VMF account can agree with naïve qualia realist on this at least: systems that are exactly functionally similar may nevertheless differ in some of their qualia concepts. Both views acknowledge a stronger sense of "translation," in which one thought is the translation of another only if it shares the very same concepts. In this sense, no one can know what it is like to be anyone else; only A can know what it is like to be A. The advantage of the VMF account is that it is able to explain this view of qualia with entirely functionalist, physicalist resources.

2.2.3 DO QUALIA EXIST?

Both the VMF account and Harman's account of qualia reject naïve qualia realism on the one hand, and eliminativism on the other. That is, both accounts of qualia are revisionist, at least in the sense of accepting NP and yet refusing to accept NE (see section 1). That is, they do not start by granting that qualia have the properties standardly believed to be had by them, and then explaining these properties in functionalist terms.

Further, as we have defined the term at the outset, Harman's account is solidly revisionist in asserting that qualia exist. But as has been hinted a few times above, the VMF account is more circumspect. Given its empirical flavor, it must be.

To understand why, it might be useful to see what goes wrong when one tries to derive an *a priori* commitment

to the existence of qualia from the VMF proposal. "On the VMF account," one might think, "the term 'qualia' refers to whatever happens to cause people to use that term. So it can't fail to refer, even if the referent is quite other than what people might think it to be. So qualia must exist."

Someone could be forgiven for understanding our proposal in this way, since our statement of what "qualia" refers to is so quick and simple. But, in fact, leaving things this way would place the bar too low for referential success. Presumably, on this simple view, "phlogiston," "witches," and "mermaids" also would refer to whatever happens to cause people to use those terms, and so phlogiston, witches, and mermaids exist, albeit in a revisionist sense[38] of the functionalist's attempt to save propositional attitudes). We do not wish to trivialize the revisionist position by adopting this simple view. Instead, we acknowledge that it is a substantive, empirical matter whether out of the possible myriad causes of "qualia" talk there is anything sufficiently unified to serve as the referent of that term (as there is not for "phlogiston," "witches," and "mermaids").[39] Further, it is not just the causes of qualia talk that play a role here, but also qualia thought, at least of the kind where one intends to employ the same concept in thought as one expresses with the word "qualia." A key *claim* of the VMF approach is that virtual machine states of a certain kind have properties that would suffice as the referents of "qualia." A key *hypothesis* of the VMF approach is that there are, in fact, such states in humans and some animals. But it is part of the VMF approach that we might discover through empirical investigation that that key hypothesis is false. Our physicalist inclinations would then, in the absence of any other acceptable account of how "qualia" could refer, push us from illusionism to hallucinationism. But such eliminativism will incur the extra demand of having to explain not only why it seemed that there were things that were ineffable, things that were intrinsic, things that were private, and things that were immediate, but also why all these seemed to be the same thing.

Compare Kevin O'Regan, who writes the following in a piece entitled "Explaining what people say about sensory qualia":

> Independent of [the debate concerning the existence of qualia] there are things people usually say about their sensory experiences that relate to the notion of qualia. People say that they cannot completely describe the "raw", basic, ultimate aspects of their sensations (e.g., the redness of red) to others (this is usually termed "ineffability"). They say that even if they cannot describe these aspects, they can be compared and contrasted (I shall say they have "structure"). And people say that there is "something it's like" to have these raw sensory experiences (they have "sensory presence"). *Whether or not qualia should be taken to exist from a philosopher's point of view*, these three things that people say about their sensory experiences need to be explained. In this chapter I show how . . . we can understand what we might mean when we say these things, *independently of whether qualia actually exist*.[40]

The inclusion of the words we have emphasized ("Whether or not qualia should be taken to exist from a philosopher's point of view" and "independently of whether qualia actually exist") makes O'Regan, to our lights, the same kind of agnostic revisionist that we are. One difference, however, is that we suspect that our account will only be fully explanatory when it reaches a certain depth of detail, and that at that point it will likely be possible to tell whether the properties of the relevant virtual machine states (if any!) are sufficiently unified to count as referents of "qualia." So we are not now, nor are we likely to ever be, in a position where we can say, "Here's an explanation of qualia, but we don't know if they exist." On the contrary, we have explained in outline how it is possible for them to exist and to play important roles in both scientific explanations and engineering designs.

In closing, we can't resist pointing out a twist that might present itself in the case in which our key claim is true, but our key hypothesis turns out to be false. That is, if we are right that properties of virtual machine components of the appropriate, unified sort are well suited to be the referent of "qualia," but we are wrong that there are such unified, suitable virtual machine components in humans (or other organisms), we could nevertheless imagine constructing an artificial agent which acquired—through evolution, learning, or design—the required unified virtual machine components. If, as we claim, such properties would likely lead such agents to develop and use the kinds of concepts we discuss above, then we might find ourselves in the awkward situation where humans do not, and yet robots do, have qualia! If the robots were philosophically sophisticated enough, some of them might even embrace doubly incorrect views of the situation, claiming that they lacked the qualia of their human forerunners because they were not biological, or because they could be completely understood in functionalist terms.

NOTES

1. Although the second author played a leading role in developing the original virtual machine functionalism account of qualia in Sloman and Chrisley, "Virtual Machines and Consciousness," the current paper is mainly the work of the first author. An unpublished document developing some of these ideas and comparing them with closely related work by Maley and Piccinini ("Get the Latest Upgrade: Functionalism 6.3.1.") is available at http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html
2. Smith, *The Problem of Perception*; emphasis added.
3. Crane and French, "The Problem of Perception"; emphasis added. On the other hand, we ourselves can only accept these definitions as they stand if virtual machine states are counted as "physical" and "ordinary" objects, a contentious view that we do not wish to defend here. A better move for our purposes would be to generalize the definitions: Illusion is any (including interoceptive) perceptual situation in which an object is actually perceived, but in which that object perceptually appears other than it really is; hallucination is an experience which seems exactly like a veridical (possibly interoceptive) perception of an object but where there is no such object there to be perceived."
4. Frankish, "Illusionism as a Theory of Consciousness," 1; emphasis in original.
5. Noë, "Is the Visual World a Grand Illusion?"
6. Pereboom, *Consciousness and the Prospects of Physicalism*, 3. As Pereboom's position has only recently come to our attention, we have not yet had a chance to analyze his insights in this area; we hope to do so on a future occasion.

7. McDermott, "AI and Consciousness," Section 3.4.
8. Sloman and Chrisley, "Virtual Machines and Consciousness."
9. Graziano, *Consciousness and the Social Brain*, 56; emphasis added.
10. Frankish, "Illusionism as a Theory of Consciousness," 3.
11. Chrisley, "Philosophical Foundations of Artificial Consciousness."
12. Turing, "Computing Machinery and Intelligence."
13. Sloman and Chrisley, "Virtual Machines and Consciousness."
14. Harman, "The Intrinsic Quality of Experience"; Harman, "Explaining an Explanatory Gap."
15. These are referred to as "architecture-based" concepts in Sloman and Chrisley, "Virtual Machines and Consciousness."
16. Kripke, *Naming and Necessity*.
17. Harman, "The Intrinsic Quality of Experience," 41.
18. A similar position is put forward in Dretske, *Naturalizing the Mind*, and Tye, *Ten Problems About Consciousness*.
19. Sloman and Chrisley, "Virtual Machines and Consciousness."
20. A notable model of how this is done and how it can go wrong (for example, in schizophrenia) is the comparator model of control; see, e.g., Frith et al., "Explaining the Symptoms of Schizophrenia: Abnormalities in the Awareness of Action."
21. Note that for argument's sake we are focusing here on instances of introspection that are themselves experiences, since those are the ones that generate this particular difficulty for the kind of "inner target" notion of introspection that the VMF account assumes. But this does not rule out the possibility of non-experiential introspection; that both kinds of introspection might be understandable in similar terms would be, to our minds, another advantage of the VMF account.
22. "Identifying" may be too strong. There are relationships of implementation or realization that are not cases of identity per se that might more accurately characterize the relationship between experiences and virtual machine states. But further consideration of these metaphysical details will have to be left to another occasion.
23. This illustration employs a familiar computational architecture for expository purposes, but it should be stressed that the notion of computational determinants of a non-occurrent virtual machine state applies much more generally. Specifically, use of this notion does not restrict us to stored program, von Neumann, serial, etc., architectures.
24. Harman, "The Intrinsic Quality of Experience," 33–34.
25. Ibid., 47.
26. Ibid.
27. Ibid.
28. Ibid., 49.
29. Although they are not in principle private or ineffable, they may, like features of a complex running software system lacking sophisticated debugging tools, be inaccessble to the program, the programmer, or anyone else, just as many of the intricate neural and electrical operations in brains are, in normal circumstances, undetectable by physiologists or physicists. That does not make them metaphysically mysterious or beyond the reach of scientific theory. Most software engineers will have had experience of making such normally inaccessible VM states and processes temporarily detectable during testing and debugging. The techniques required are very different from those required for detecting physical and chemical states and processes. In both cases there is always a danger that the observation processes may alter what is observed. (This has nothing to do with quantum mechanics.)
30. Especially if they are unfamiliar with the kinds of concepts one typically acquires from first-hand experience of developing, testing, and debugging complex running virtual machinery.
31. The rest of this paragraph, and the next two, reproduce page 167–68 of Sloman and Chrisley, "Virtual Machines and Consciousness," nearly verbatim.
32. Campbell, *Past, Space, and Self*, 43.
33. Lycan, *Consciousness and Experience*.
34. Jackson, "Epiphenomenal Qualia."
35. Mandik, "Mental Representation and the Subjectivity of Consciousness," 185.
36. Harman, "Explaining an Explanatory Gap," 2.
37. Ibid., 3.
38. Compare the criticism in Churchland, "Eliminative Materialism and the Propositional Attitudes," 81.
39. See Cussins, "Nonconceptual Content and the Elimination of Misconceived Composites!" for one account of what "sufficiently unified" might amount to.
40. O'Regan, "Explaining What People Say about Sensory Qualia," 31–32.

REFERENCES

Blackmore, S. "The Grand Illusion: Why Consciousness Exists Only When You Look For It." *New Scientist* June 22, 2002, pp. 26–29.

Campbell, J. *Past, Space, and Self*. Cambridge/London: The MIT Press, 1994.

Carruthers, P. "Précis of *Phenomenal Consciousness*," 2001. http://www.swif.uniba.it/lei/mind/forums/forum2.htm, accessed January 5, 2008.

Chrisley, R. "Philosophical Foundations of Artificial Consciousness." *Artificial Intelligence in Medicine* 44, no. 2 (2008): 119–37.

Churchland, P. "Eliminative Materialism and the Propositional Attitudes." *The Journal of Philosophy* 78, no. 2 (1981): 67–90.

Cussins, A. "Nonconceptual Content and the Elimination of Misconceived Composites!" *Mind and Language* 8 (1993): 234–52. doi:10.1111/j.1468-0017.1993.tb00283.x

Crane, T. and French, C. "The Problem of Perception." *The Stanford Encyclopedia of Philosophy*, spring 2016 edition, edited by Edward N. Zalta. http://plato.stanford.edu/archives/spr2016/entries/perception-problem

Dennett, D. "Quining Qualia." In *Consciousness in Modern Science*, edited by A. Marcel and E. Bisiach. Oxford University Press, 1988.

———. "Who's on First? Heterophenomenology Explained." *Journal of Consciousness Studies* 10 (2003): 19–30.

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

Frankish, K. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies*, forthcoming.

Frith, C., S. Blakemore, and D. Wolpert. "Explaining the Symptoms of Schizophrenia: Abnormalities in the Awareness of Action." *Brain Research Review* 31 (2000): 357–63.

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

Harman, G. "The Intrinsic Quality of Experience." *Philosophical Perspectives* 4 (1990): 31–52.

———. "Explaining an Explanatory Gap." *The American Philosophy Association Newsletter on Philosophy and Computers* 6, no. 2 (2007): 2–3.

Jackson, F. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (1982): 127–36.

Kripke, S. *Naming and Necessity*. Cambridge: Harvard University Press, 1980.

Lycan, W. "Inverted Spectrum." *Ratio* 15 (1973): 315–19.

———. *Consciousness and Experience*. Cambridge, MA: The MIT Press, 1996.

Maley, C., and G. Piccinini. (2013) "Get the Latest Upgrade: Functionalism 6.3.1." *Philosophia Scientiae*, 17, no. 2 (2013): 1–15. http://poincare.univ-nancy2.fr/PhilosophiaScientiae/

Mandik, P. "Mental Representation and the Subjectivity of Consciousness." *Philosophical Psychology* 14, no. 2 (2001): 179–202.

Noë, A. "Is the Visual World a Grand Illusion?" *Journal of Consciousness Studies* 9, nos. 5-6 (2002): 1–12.

McDermott, D. "AI and Consciousness." In *The Cambridge Handbook of*

Consciousness, edited by P. Zelazo, M. Moscovitch, and E. Thompson, 117–50. New York, NY: Cambridge University Press, 2007.

O'Regan, K. "Explaining What People Say about Sensory Qualia." In *Perception, Action, and Consciousness*, edited by N. Gangopadhay, M. Madary, and F. Spicer, 31–50. Oxford University Press, 2010.

Pereboom, D. *Consciousness and the Prospects of Physicalism*. Oxford University Press, 2011.

Schwitzgebel, E. (2014) "Introspection." *The Stanford Encyclopedia of Philosophy*, summer 2014 edition, edited by Edward N. Zalta. http://plato.stanford.edu/archives/sum2014/entries/introspection

Shoemaker, S. "The Inverted Spectrum." *Journal of Philosophy* 79 (1982): 357–81.

———. "Self-Knowledge and 'Inner Sense': Lecture I: The Object Perception Model." *Philosophy and Phenomenological Research* 54, no. 2 (1994): 249–69.

Sloman, A., and R. Chrisley. "Virtual Machines and Consciousness." *Journal of Consciousness Studies* 10 (2003): 113–72.

Smith, A. *The Problem of Perception*. Cambridge, MA: Harvard University Press, 2002.

Turing, A. "Computing Machinery and Intelligence." *Mind* 59 (1950): 433–60.

Tye, M. *Ten Problems About Consciousness*. Cambridge, MA: The MIT Press, 1995.

———. "Qualia." *The Stanford Encyclopedia of Philosophy*, Fall 2015 edition, edited by Edward N. Zalta. http://plato.stanford.edu/archives/fall2015/entries/qualia/

From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness, Part 1

Jeffrey White

KOREAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY (KAIST) COMPUTATIONAL NEUROSYSTEM LABORATORY, DEPARTMENT OF ELECTRICAL ENGINEERING, DRWHITE@KAIST.AC.KR

Jun Tani

KOREAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY (KAIST), DEPARTMENT OF ELECTRICAL ENGINEERING

ABSTRACT

Direct neurological and especially imaging-driven investigations into the structures essential to naturally occurring cognitive systems in their development and operation have motivated broadening interest in the potential for artificial consciousness modeled on these systems. This first paper in a series of three begins with a brief review of Boltuc's (2009) "brain-based" thesis on the prospect of artificial consciousness, focusing on his formulation of h-consciousness. We then explore some of the implications of brain research on the structure of consciousness, finding limitations in biological approaches to the study of consciousness. Looking past these limitations, we introduce research in artificial consciousness designed to test for the emergence of consciousness, a phenomenon beyond the purview of the study of existing biological systems.

SECTION 1: INTRODUCTION

Nature seems here eternally to impose a singular condition, that the more one gains in intelligence the more one loses in instinct. Does this bring gain or loss?

– Julian Offray de La Mettrie[1]

The following paper is the first of three. It sets out the case for research in artificial consciousness, arguing that studies in artificial systems are a necessary complement to research into biological systems due both to the nature of artificial systems as well as the limitations inherent in studies of biological systems. First, it briefly introduces Piotr Boltuc's "naturalistic non-reductionist" account of consciousness which holds that "first person consciousness is not reducible to material phenomena, but that it is at the same time fully explainable by such phenomena."[2] Then, the second and third sections of this paper explore some of the implications of studies into biological consciousness, one of which being that the "pure" subjectivity that is the object of some philosophical discourse is quickly occluded by concomitant processes and overlapping networks. Through the discussion, Boltuc's originally clear assay gives rise to two more complex types of consciousness, **most-consciousness** and **myth-consciousness**, both apparently necessary and not accidental aspects of human cognitive agency. We find a complimentary account in recent work from Thomas Fuchs, and here are met with practical limits to consciousness research in biological systems. In the third section, we follow Edelman and Baars in looking directly at research into artificial consciousness as a way past these limitations. Finally, the fourth section quickly reviews a series of experiments establishing the emergence of a minimal self-consciousness in lead up to the second paper in this series, which reviews this group's most recent work on freewill.

Concerning artificial consciousness, Boltuc has issued a positive thesis. He is confident that artificial consciousness is possible when the material nature of biological cognition is better understood. "Machines can be conscious like any organism can."[3] He offers an analysis of consciousness into three forms, functional, phenomenal and h-consciousness ("hard"), and he raises questions about a locus of consciousness based on existing biological systems.

On Boltuc's estimation, robots are already what he calls "functionally" conscious. Through their normal function, "they can perform many thinking tasks comparable, or superior, to humans, though by other means."[4] "Thinking" for Boltuc is simple enough, being "any kind of information processing that increases inductive probability of arriving at a correct result"[5]—i.e., error correction. So, thinking is integral to learning. Phenomenal consciousness is more complex, and at the center of what Boltuc takes to be "*the* most important, but somewhat neglected, philosophical issue in machine consciousness today", that "every function attributed to p-consciousness could, in principle, be played by an AI mechanism using some sort of functional mechanism, only."[6] That this is not yet the case is due specifically to the lack of an adequate "generator of consciousness" the functions of which, once understood adequately, will be able to be engineered.[7]