

Philosophy and Computers



FALL 2016

VOLUME 16 | NUMBER 1

FROM THE EDITOR

Peter Boltuc

FROM THE CHAIR

Marcello Guarini

MIND ROBOTICS

Ron Chrisley and Aaron Sloman

Functionalism, Revisionism, and Qualia

Jeffrey White and Jun Tani

*From Biological to Synthetic
Neurorobotics Approaches to
Understanding the Structure Essential to
Consciousness (I)*

Riccardo Manzotti

*Objectbound: A Mind-Object Identity
Theory*

Jordi Vallverdú and Max Talanov

*On Importance of Life and Death for
Artificial Intelligence Creatures*

ETHICS ROBOTICS

Marcello Guarini

Carebots and the Ties That Bind

Sean Welsh

*Formalizing Hard Moral Choices in
Artificial Intelligence*

Scott Forschler

*The Nature of Avatars: A Response to
Roxanne Kurtz's "My Avatar, My Choice"*

Roxanne Kurtz

IRL Rejoinder to Scott Forschler

CALL FOR PAPERS



APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR

VOLUME 16 | NUMBER 1 | FALL 2016

FROM THE EDITOR

Peter Boltuc

UNIVERSITY OF ILLINOIS, SPRINGFIELD

Marcello Guarini is the new chair of the APA Committee on Philosophy and Computers. **Warm welcome, Marcello!** In his note from the chair, Marcello presents an ambitious project of bringing the charges of the committee up to date with the actual role it plays; much has changed in the area of philosophy and computers over the last two decades. Marcello's article on artificial companions opens Part II of the current issue.

Ron Chrisley and Aaron Sloman open Part I of this issue with their article "Functionalism, Revisionism, and Qualia." Chrisley and Sloman discuss revisionism about qualia—the view that tries to navigate between naïve qualia realism and reductive eliminativism. The authors discuss the relevance of their approach to AI. They also relate to the works they view as following the main tenets of revisionism about qualia. This includes Gilbert Harman's version of functionalism, discussed in much detail (including Harman's article "Explaining the Explanatory Gap," published in the spring 2007 issue of this newsletter) and also the psychomotoric approach to qualia by Kevin O'Regan.

The first four contributions in this issue relate to topics in philosophy of mind and consciousness relevant for robotics; this includes a philosophical cartoon by Riccardo Manzotti, which I decided to place next since it is a perfect follow-up to the paper by Chrisley and Sloman. Manzotti advocates what he calls relativism between objective and subjective properties as a way to bridge the mind body gap (or the Hard Problem of Consciousness). While I would use the word "relationism" rather than "relativism," the project is presented persuasively (I am currently under the influence of O'Regan's psychomotoric view, which seems to give a simpler explanation that goes in a similar direction). The article by J. White and J. Tani also deals with the Hard Problem. The authors discuss the use in human and robot cognitive architectures of the notion of h-consciousness and add their own notions of most-consciousness and myth-consciousness (based on Fuchs 2016). The essay moves from analytical to phenomenological considerations.

The last section (Section 4), based on the work of J. Tani, will be of special interest to those into neurorobotics. The focus on the relevance of time for person-creation and the "motivational potential arising as horizons of anticipation are projected due normal adolescent neural development," is a

perfect *segue* into the article by J. Vallverdú and M. Talanov. Their article argues that humanoid robots should actually be aging. Since human life has a limited time horizon, a humanoid robot without having—and understanding that it has—such time horizon would not be humanoid in an important sense. The authors present quite controversial points targeted at gerontocracy (those points may reflect the experiences in some non-democratic countries, when death of the leader, not the elections, provide the best opportunity for change)—perhaps we should not preclude a longitudinal, perhaps modular, neural network able to learn and adapt. It could be designed in such a way that it increases in the relevant functionalities over millennia; let us call it an aleatory network. Designing one may be a worthy challenge.

The second part of the issue is devoted to robot ethics. We open with the article by Marcello Guarini, which pertains to the very timely issue of artificial companions. The author builds an argument trying to hit the golden mean between the skepticism or even pessimism of Sherry Turkle and Luciano Floridi's well-known optimism. Guarini focuses on the benefits for our life and character that come from caring for our significant others, which may be undermined by extensive use of advanced robotic caregivers of the future. Next, Sean Welsh, our first contributor from New Zealand, presents some philosophical ideas towards formalization of the hard problems of ethics, such as the trolley cases. Welsh follows on the work of Jim Moor and others who use the need for formal solution in building *ethics sub-routines* for self-guiding robots as an opportunity to clarify our best intuitions in applied ethics. We close with a discussion between Scott Forshler and Roxanne Kurtz about the moral and ontological status of avatars, which pertains to Roxanne's paper presented at the APA and published in an earlier issue of this newsletter.

FROM THE CHAIR

Marcello Guarini

UNIVERSITY OF WINDSOR

As the new chair of the APA Committee on Philosophy and Computers, I thank Thomas Powers for his leadership of the committee over the past few years. Tom's selfless and consultative approach has made it a pleasure for all of us to work with him. The committee, and the community of scholars served by it, are better off as a result of Tom's efforts.

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 Sloman¹
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 consciousness⁵ 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 embraced⁷ 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 brains¹²).

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 experience²¹ (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) consciousness, 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 identifying²² 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 response³³ 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 sense³⁸ 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 inaccessible 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. Gangopadhyay, 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 Mettrie¹

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 essay 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

Boltuc analyzes p-consciousness into subcategories, the “broad” also “functional” sense indicative of “first-person functional consciousness” including direct perception, and the “narrow” “non-functional” sense indicative of the “mineness” that characterizes human-like “hard” “h-consciousness”. H-consciousness is the focus of Boltuc’s engineering thesis⁸ because it represents the “awareness” of being, “the locus of first-person experiences”, and he argues that without this awareness “there is nothing that it is like to be that robot.”⁹ Important to Boltuc’s analysis here is his distinction between subject and object. A subject is ultimately a non-object, and an object a non-subject. For Boltuc, this constitutes the simplest ontology, and helps to further clarify the special nature of h-consciousness. One way to understand the first-person perspective is as that “subjective perspective from which one performs a certain function (e.g., the perspective from which one makes a picture)” but another way is “the very stream of awareness that a conscious individual has.”¹⁰ One is “inside” and the other, the former, remains a third-person perspective on the first-person perspective. For Boltuc, this distinction underscores the difference between a proto-cognitive system like a camera, or a robot with some minimal degree of “consciousness”, and the different case that is h-consciousness. He labels it the “Is anybody home?” problem essential to “systems with their own locus of awareness.”¹¹

The “is anybody home” problem has to do with feeling the difference between before and after states consequent on thinking actions, as an agent experiences and necessarily (perhaps permanently) embodies what at least seems to follow from conscious phenomena. Being able to answer the “Is anybody home?” question is the reason that h-consciousness is “at least a condition of one’s status as a moral patient strictly understood.”¹² And, given also “a strong, plausible tendency to view moral value as dependent on first person awareness (h-consciousness)”, the ability to answer this question carries “strong implications for ethics and in particular for the relative moral standing of robots, as they are now, and animals (including humans).”¹³ After all, Boltuc is not ready to afford moral status to entities simply because they are h-conscious, e.g., “rats.”¹⁴ More seems to be needed, and we will begin to look into what this more might amount to in the next section.

Finally, Boltuc argues that an artificial consciousness is unlikely to emerge as an aspect of a computer “program”. On his assay, a program can model complex biological systems, but that “they are not those systems” and therefore “it is very unlikely that h-consciousness is merely a feature of a program.”¹⁵ His advice is to pursue inquiries into biological systems first, understand for example what differentiates human cognition from that of a rat, and from this end “try to build a generator of consciousness in some other, inorganic or organic, matter or, if possible, find them in some already existing systems” i.e., as “generated in human and other animal brains.”¹⁶ And, this is where we are left, with the challenge to both conceive of how h-consciousness can be fully explained in material terms through an understanding first of how h-consciousness is “generated” in available biological systems and then, with this understanding, to “engineer” it. This is Boltuc’s non-reductive naturalistic thesis.

There are some immediate problems with such a proposal. For one thing, success is effectively impossible to confirm due to the fact that the “verification of results” is confounded given the privileged access that characterizes h-consciousness as the “mineness” of experience. The work now is to account for this mineness in the most direct way. Consider Michel Bitbol’s view.¹⁷ On Bitbol’s estimation, subjectivity is not something extra, it is essential to cognition, for consciousness, yet so is objectivity and the result is a necessary “dance” between first and third person perspectives in the representation of consciousness as a “stabilized and intersubjectively shared structural residue.”¹⁸ Problems with privileged access to subjective states and the problem of other minds remain imperfectly resolved, but this is the nature of conscious systems and to be expected. Such is not the nature of artificial conscious systems, however, and from them we may form different expectations.

The hurdle of privileged access may be overcome with perfect information about the dynamic structure of a cognitive system ready at hand, a perspective not afforded human observers of natural cognitive systems *in situ* such as those which are Bitbol’s and Boltuc’s main concern. This potential is afforded, however, by artificial systems as we shall see in greater detail going forward. First, we must further establish the limits of the use of biological models in the search for a formal structure of consciousness.

SECTION 2: TEMPORALITY

At issue is the potential for conscious machines, specifically artificial systems with a sense of ownership over their actions and intended ends. Piotr Boltuc has advised that research on the hard problem of engineering artificial consciousness should focus on the structure of “existing systems” in order to understand consciousness as it may be made to exist in non-human artificial agents, specifically through engineering “projectors of first-person awareness.”¹⁹ Consider this fact in approaching the problem of consciousness as posed in Chalmers’ zombie thought experiment.²⁰ A focus on structural isomorphism doesn’t seem very promising in solving the hard problem of consciousness in zombies, as these are structurally identical with existing conscious systems. On the form of Chalmers’ thought experiment, consciousness must be something other than structural isomorphism at the finest grain of material assay.

Chalmers’ zombies help to spotlight the fact that with every reduction of consciousness into material nature there remains the question, what is missing in a zombie equivalent. This is the “hard” problem of consciousness. There are different ways of trying to zero out the debt that remains on the “full” explanation of consciousness in purely material terms.²¹ One may confess to being a zombie. One may posit the existence of a locus of the feeling of mineness of consciousness, typically some *organelle* and corresponding operations within a biological brain without which consciousness in whatever form is impossible, which is the general direction recommended on Boltuc’s thesis as well.²² Some lines of inquiry isolate consciousness to networks of activity at the center of which is a hub of activity in the thalamus, with ongoing work in the structure

internal to the thalamus and how this anatomy correlates with consciousness.²³ This is not a unique view; it is not unpopular and not new given the longstanding recognition of the thalamus as a special hub of neural activity central to consciousness.²⁴

However, if we are to look at isolating a distinct region of central neural activity as the locus of h-consciousness in particular, then the thalamus may not be the best candidate area. After all, Boltuc's engineering thesis merely advises that any effort at artificial consciousness should aim to recapitulate something performing as a "projector" or "generator" of consciousness, and this role might be played by a number of candidate systems. One possibility is the reticular activating system, for example. The reticular formation sits at the confluence of the internal environment of the central neural system above it and the external perceptual reality as mediated by the body system below. And, its role in the "projection" of consciousness is well-known, for example as set out by Parvisi and Damasio who argue that consciousness arises when an organism is able to "internally construct and internally exhibit a specific kind of wordless knowledge" that "the organism has been changed by an object ... along with the salient enhancement of the object image caused by attention being allocated to it."²⁵ Consciousness then arises as the agent adjusts to the experience, a process enabled by the reticular formation and carried forward by the reticular activating system.²⁶

That said, many regions above the reticular formation seem to be even more important to the sense of "mineness" characterizing h-consciousness in particular. When searching for the locus of the feeling of what it is to be "me" rather than another subject, consider the ventromedial prefrontal cortex which—as Bechara, Damasio and Damasio describe²⁷—"links" particular perceptions with established emotional valences. Moreover, these associations are then modulated, especially reinforced post-choice in the reduction of internal inconsistencies ("cognitive dissonance") and without awareness.²⁸ Folded into the discussion thus far, it is difficult to imagine what could be more "mine" than the surprising feeling of an unexpected adjustment to a disposition to act. And yet, the vmPFC is involved in processing specific to other essential ingredients of the mineness of experience, as well.

If anything were more mine than the felt update on prior embodied yet hidden preferences, then it may be one's anticipations of a personal future. Damage to the ventromedial region correlates with the inability to take up anticipatory emotional states as evidenced by skin conductance on the presentation of a decision situation, with subjects optimizing for short rather than long term rewards, "oblivious to the future."²⁹ Other research has demonstrated that reduction in activity of the vmPFC correlates with reduced predictive capacity due to the fact that the vmPFC enacts processes that effectively populate possible futures from the first-person perspective such that a failure in predictive capacity ultimately derives from the "failure to think self-referentially about our future selves."³⁰

In short, due to activity in the vmPFC, we may say that a human cognitive agent "has a future", and moreover we

may say that this future is essentially social with deep moral implications for a biologically realistic account of h-consciousness, as well. The vmPFC is implicated in "empathic decision making" which involves making decisions in order to optimize another's future well-being.³¹ Accordingly, vmPFC damage has been associated with impaired moral emotions such as empathy central to morality and implicated in moral judgment.³² The right vmPFC especially is implicated in empathy, with damage to this area resulting in, among other things, reduced moral sensitivity to situations involving perceived injustice.³³ The central thesis here is that evolved biological drives result in "moral emotions" that in the vmPFC automatically conjoin self and other interests in constraining possible futures towards which cognitive agency is then exercised. The result is the creation of joint attention and "intersubjective space" as the default form of future into which a self is projected in part through vmPFC processes (and echoing Bibbol in an interesting way). Taken as a whole, this research affords insight into the essentially social nature of human cognition due the essential social nature of the generation of the possible future self through activity in the vmPFC in particular.³⁴ Human cognitive agency is social agency, simply put.³⁵

Here, we find a locus of activity contributing to the mineness of h-consciousness that is at the same time essentially social and also temporal, outstripping Boltuc's original analysis of h-consciousness as pure subjectivity. And as we explore the implications of this activity, the original analytic sense of zombie has also mutated into something more, something closer to actual human beings, perhaps moral zombies instead. After all, Boltuc intends merely that h-consciousness is a necessary but insufficient condition for being a moral patient, as a "locus of awareness" characteristic of first-person experience. Yet, the mineness characteristic of h-consciousness as we have been developing it, in consideration of neural processes and how these are bundled, reveals the essentially social and temporal dimensions of what we may call "**most-consciousness**" (mine other self temporal) instead of merely h-conscious in order to differentiate from Boltuc's analysis.

vmPFC processing makes it a prime candidate as a locus of most-consciousness for an essentially social cognitive agent, perhaps especially when integrated with the dmPFC into the entire mPFC.³⁶ After all, if anything were more characteristic of the "mineness" of experience than the surprising adjustment to erstwhile hidden preferences and redirection of one's future project self, it may be the empathy opened to others, directly allowing one's self and its projected future to be emotionally transformed through moral perspective taking. With this in mind, then, an easy answer to "What is the most-zombie missing?" is "a future" or perhaps "a future with friends in it" or perhaps just as well, a vmPFC. Finally, if we accept as essential the relationship between a project future central to mineness and moral agency exercised toward this internally constructed end then it stands to reason that one way to engineer a zombie without a sense of ownership of its own agency is to somehow interfere with the function of the thalamus, or with the vmPFC, to take its present, or its future, or both.³⁷

But, what about its past? Time consciousness involves not only future and present, but also past. Nothing may be more “mine” than my own future, and how I feel about it, except perhaps my own past, how this brings me into the present and disposes³⁸ me toward some futures rather than others. Without a past, one may be sensitive to changes without recognizing the difference between before and after as if on a perpetual roller coaster with no time to think. Likewise, we may imagine that zombies may be without pasts, without memories, without the mineness that characterizes most-consciousness.

Memory formation is thought to depend mostly on another area of the brain, the hippocampus, and interfering with the function of the hippocampus can result in something like a zombie. One interesting and more or less common loss of most-consciousness corresponds with the loss of memory in an alcohol blackout. The mineness of consciousness is lost along with the feeling of before and after. During blackouts, affected individuals often execute complex action routines including speech and the use of symbols within noisy and even dangerous environments while being left with often spotty memories which seem to indicate that p-consciousness was in some limited way present, but lost.³⁹ Of the rest, there is no sense of mine-ness. There is no memory. Something here is missing, Boltuc’s “very stream of awareness” is interrupted, and this is what makes an alcohol blackout like being a zombie.⁴⁰

When we think most broadly about the constituents of a unique self, especially about what is unique to this agent as opposed to any other, we might be drawn to the notion of memory. The vmPFC is necessary, and the thalamus, certainly, and all are structurally and functionally unique to each subject at the finest grains of analysis, but without a memory of how one used to feel about something, before that preference changed, then the “mineness” characteristic of most-consciousness is also impossible.⁴¹ In a way, then, the hippocampus seems to be a good location on which to focus if one were intent on the creation of most-zombies, i.e., beings exactly like us but without most-consciousness. However, it is not a difference in neural anatomy that makes the difference here. Rather, it is the presence of magnesium ions within an ion channel that modulates memory formation. Blackouts happen when Mg⁺⁺ doesn’t get into the channel to block the influx of ions because without this plug, impaired memorization and long-term brain damage result.⁴² Thus, to the question “What is the zombie missing?” one may answer “Magnesium ions in receptor channels modulating NMDA receptor function” rather than name any neuroanatomical organelle.

Moreover, if we can imagine a drug which keeps these channels open to an influx of ions that results in the burnout of memory formation – perhaps permanently and reliably given certain selective stimuli – then we can imagine the purposeful creation of zombies which are potentially selectively incapable of consciousness of certain things and relations. The field of perception can be stabilized by unconscious processes, and so a stable subject is estimable, but in fact the feeling of mineness about one’s own direct experiences would be absent without a sense of change at least seemingly dependent on conscious processes

and moderated by past conditions which are absent, on this rather soft zombie example.⁴³ Without this before and after, we are left with a mindless doppelganger due an alcohol blackout, a being with most-conscious potential but without the memory that partly constitutes the sense of self about which most-consciousness is concerned.

Even this construct does not help us to solve the hard problem as originally formulated in Chalmers (1996) because at the finest grains of analysis it is not structurally identical with a non-zombie. Due to the presence or absence of ions and other neurotransmitters, molecular conformations change. Potentials for development change. Futures change, and even disappear. But, that doesn’t mean that we can’t get clearer on the relationship between consciousness and memory by holding the alcohol blackout alongside the zombie model. To be a zombie requires that the subject first have the potential for most-consciousness, and then to be denied its realization. The question is after all in the form of a “What is missing?” And, memory certainly qualifies as a natural non-reductionist candidate, fully explainable but not fully reducible to material description, after all being dependent on context and interpreted for others including one’s self in reflective inner discourse. At the same time, there is a strong association between memory and moral agency as illustrated in the fact that human beings may be exempted from misdeeds performed during blackout states that would otherwise invite greater sanction.⁴⁴

The flip-side of the problem of other minds is the issue of accounting for one’s own. How much must be accounted for, and what is the best way to do it? How much memory does a cognitive agent need in order to be conscious in a morally relevant way, not be snuffed out as a nuisance? More than a rat? What kinds of memories are necessary? What kind of future is necessary? And, if these are all necessary, then isn’t the hippocampus also necessary along with the ventromedial and the thalamus? Where with a biological model of consciousness must we stop for an adequate account of consciousness? With the brain of the agent? Its skin? The systems in terms of which it is embedded?

Here, especially, we can see the role for symbolic expression in the construction of narratives that make conscious exposition possible. Symbols help us to remember. And they also help us to project. From ink and paper to the printing press, the first popular fictions were psychological self-reports.⁴⁵ With these narratives as subjective, first-person anticipatory and regretful accounts of life from the inside-out so to speak, there is the modern sense that what is important is not determined and the past perhaps best left behind, with the future open and at least potentially within control, the modern project which so given represents simply a ubiquitous aspiration intersubjectively distilled.⁴⁶ We will have something to say, in the third paper, about the motivational potential arising as horizons of anticipation are projected due normal adolescent neural development.

In the end, if articulating artificial consciousness means simulating all of this complexity in a computational medium, e.g., artificial cognitive agents which write papers for publication on the prospects of artificial cognition, then we may well have before us an impossible task.

SECTION 3: MYTH-CONSCIOUSNESS

We have been sorting out how to understand biological models of consciousness in a way which affords a neat view on especially the “mineness” of consciousness which we have since developed from an analytic shell into a biologically more realistic sense of most-consciousness. But what is the relationship between consciousness and cognition generally speaking? In his exposition, Boltuc defines “cognition” as “interactions of a system with an environment” and importantly he adds a requirement, that this system must “involve structural retention of some pieces of information”. On this “somewhat zoocentric” account, for both biological and robotic “organisms” cognition can be construed without “reference to consciousness, as processing of sensory input and its assimilation into existing schemes” as the agent “gains knowledge” and “becomes aware” and then “uses that knowledge for comprehension and problem solving.”⁴⁷ Cognition is very much like “thinking” which on Boltuc’s recipe works essentially to solve problems. The essential difference is that structural retention alone such as that involved in “learning” remains “short of consciousness” along with “sleepwalking” until “down towards simpler organisms”, e.g., “roaches,” questions for example about moral status due to consciousness become meaningless “since we seem to have no reason to presume consciousness apart from those [unconscious cognitive] processes.”⁴⁸ Thus, Boltuc offers a definition of consciousness in relation to cognition – that it is a “special instance” of cognition, which cannot be reduced to simple cognition, yet which operates as an extension of embodied cognitive agency, the exercise of which ideally opens further opportunities for continued cognition, i.e., as the agent becomes aware of problems worth solving.

Given that problems facing agents may arise in all modes of said agency, any realistic account of the “mineness” of consciousness is unlikely to limit itself to any single region or sub-process within the brain. Is it possible, then, that we may account for consciousness in terms of more distributed neural systems? Consider Edelman, Gally and Baars on what consciousness requires:

Consciousness consists of a stream of unified mental constructs that arise spontaneously from a material structure, the Dynamic Core in the brain. Consciousness is a concomitant of dynamic patterns of reentrant signaling within complex, widely dispersed, interconnected neural networks constituting a Global Workspace.⁴⁹

“Unified” phenomena become so by the harmonic coordination of networks via the “Dynamic Core” on the thesis that consciousness arises due to cortical processes as they re-enter the thalamus from various regions. As we have seen, the thalamus is the nexus of “reentrant signaling” produced by the complex, widely dispersed, interconnected neural networks which constitute Baar’s (1997) “Global Workspace” some of which is engaged within any given task environment.⁵⁰

Dynamic core models clarify a number of things. For example, the subjective sense that consciousness is

something over and above simple cognition is reinforced in the anatomy of thalamocortical loops as more connections develop between the thalamus and the frontal areas of the cortex in the direction of the thalamus than the other way around. And, on this more systematic view, we find that consciousness arises in the synchronization of distributed operations rather than in virtue of one class of cells within one sub-region of the central nervous system. That said, though such an account may tell us why a thalamus is necessary for consciousness in biological systems like ours, it does not tell us in what form it may be essential to conscious systems in general.⁵¹

Recognizing such limitations in biologically reductive approaches to consciousness, Edelman, Gally and Baars recommend that “A theory of human consciousness... must rest on a more global theory of how vertebrate brains are organized to yield function.”⁵² And this means exactly looking past the structure internal to the brain itself for the influences shaping central neural system organization.

In this light, consider Thomas Fuchs’ effort at understanding cognitive agency as enabled deeply embodied material memorization, with the agent as a whole its own record, and this again only significant in light of an agent’s projected and anticipated future selves.⁵³ This is a “more global” account than those reviewed so far, as it begins with autopoietical self-organization and identifies consciousness with processes set on maintaining the integrity of the organism in the face of disintegrative change. Fuchs sees cognitive agency as “integrative” embodiment, with memory distributed both within the brain and also in the material processes of the distributed body system as it internalizes the world in its interactions. Consciousness, thus, emerges in the “diachronic unity” of cognitive agency, as the agent sets out and feels its distance from more or less ideal situations with this proto-natural inclination ultimately shaping how brains and bodies are organized to yield function.

“The systemic unity of the organism thus becomes the precondition of the unity of self-experience”⁵⁴ as the “diachronic unity of consciousness” is formed by “a self-referential process in which each succeeding moment implies an awareness of the next-to-come and the just-past” which results in “a pre-reflective self-awareness”⁵⁵ of the imminence of these instances as they are to be embodied. In accounting for this process, Fuchs stresses the role of self-organization of an organism individuated not by accident but due to its embodied nature and by way of which there arises the experience of “continuity of the self from a first-person perspective” the “pre-reflective feeling of sameness or a felt constancy of subjectivity” to which one awakens prior to any remembering or reconstruction of an object self.⁵⁶ And, he finds the substrate of this felt sameness in the concept of “bodily existence” itself.

Bodily existence is characterized first of all by the “diachronic coherence of a basic bodily self” and this coherence specifies only “an abstract identity or sameness . . . but no qualitative identity,” i.e., it is purely formal. It tells us nothing of “the sort of persons that we are” and it is for this same reason that it is unrealistic, neglecting

the fact that "all enactments of life are integrated into the memory of the body, and here they remain preserved as experiences, dispositions, inclinations and skills."⁵⁷ In other words, on Fuchs' account, consciousness is always and already of deeply embodied material memory with this record also establishing implicit valuations on experience that are more or less malleable (e.g., in the case of an octopus, not so much). In so far as this embodied agency is furthered or hindered, healed or injured things are good or bad, and the body as memory is the record of this status quo as well as how to deal with it.⁵⁸ "Body memory is thus the ensemble of all habits and capacities at our disposal."⁵⁹

Rather than looking for a generator of conscious phenomena, Fuchs finds the grounds of consciousness in the temporal structure of consciousness, the binding together not of subjective with objective points of view, but rather future with past and not within a self as a separable process, but constitutive of self (and likely demanding a "multi-layered" understanding of self⁶⁰). Working from a Husserlian analytic, Fuchs writes that "the stream of experience as a continuous synthesis of what is not-yet, what is now, and what is no-longer" constitutes the "diachronic unity of consciousness" as a "self-referential process in which each succeeding moment implies an awareness of the next-to-come and the just-past" resulting in, again, the "pre-reflective self-awareness"⁶¹ that is the also the target of Boltuc's h-consciousness. However, it is only when coupled with the deep material memory of embodied cognition that this "pure" subjectivity takes on its unique character, mine own such that "without its embedding in the continuity of pre-reflective bodily existence" the mineness of consciousness disappears and "the narrative self and its memories remain but [as] a story that we tell about an alien person."⁶²

It is not simply a matter of occupying a position in a course of historical evolution that is at issue, here. Rather, the capacity for the subject to employ embodied resources in the direction of this history is a difference that is worthy of distinction from the simple model of most-consciousness that we have developed thus far. Consider here why Edelman, Baars and Seth hold forth for the necessity of narrative facilitated by language in the exposition of the "mineness" characteristic of human consciousness as it facilitates the detachment of the subject from the feeling of being its self in the present.⁶³ This same capacity allows human beings to represent for others similarly embodied the series of conformations undergone in a felt, embodied transformation from one situation to another, e.g., we can learn from others' self-reported experience, and reflect on our own in the same way. On the other hand, the authors do not find this capacity in octopi as they appear unable to adapt neural structures driving goal achievement in recognizable response to contextual cues in a laboratory environment, so seemingly making any narrative progress beyond simple evolutionary forces impossible and any question about how octopi might communicate changes to others moot (for example, through skin color changes). As this case illustrates, there is a distinction to be made between a cognitive agent acting within the space of its evolution, determined by its inherited form, and a cognitive agent with the capacity to *make* this history both through symbolically represented narrative exposition as well as

through self-directed self-change, becoming the agent required for the execution of some action or other the necessity of which arising first in the subjective projection of possible future self-situations of and for that very agent. This new form we may call "**myth-consciousness.**" This distinction between what we have been calling "most-" and "myth-" consciousness is worth some attention.⁶⁴ The latter's important role comes largely from the fact that cognitive agents more or less embody the vastness of history and its determinations.⁶⁵

Myth-consciousness retains the mineness and essentially moral social temporal nature of most-consciousness, but recognizes that as embodiment is the medium of memory, and that as embodied memory is history, then the nature of (human-like) consciousness is essentially historical as well. From this point of view, "Pure consciousness without a subjective body is a dualistic abstraction which forgets that all thinking owes its emergence to the preceding process of life."⁶⁶ These life processes are the assemblage of continuously unfolding problem solving routines unique to a uniquely historically situated, uniquely materially embodied social cognitive agent, i.e., Boltuc's "thinking."

Fuchs writes that "through our habits, we inhabit the world"⁶⁷ with this habitation constraining focal attention, while also cementing the agent into the landscape of living history which is its focus. Thus we may stress that this is a circular process. The world informs (as in "puts the form into") our habits, and through these habits we change the world that again informs our habits upon which we then act.

Emphasizing the historical depth of embodied memory, Fuchs points to the difference between traumatic and more everyday memories with interesting implications for our understanding of myth-consciousness, as well. In illustration of the way that trauma affects embodied memory, Fuchs quotes Aahron Applefield who survived as a refugee from Ukraine under fascism during the last World War: "The cells of my body apparently remember more than my mind which is supposed to remember."⁶⁸ And, Fuchs is right to draw attention to the difference in deep memory of traumatic versus everyday events. Jovasevic et al. have shown in a mouse model two pathways for memory encoding active in the hippocampus, one which distributes memory to higher level cortical regions, and another which passes traumatic memories to sub-cortical systems essentially outside the reach of conscious introspection.⁶⁹ Surprisingly, the same GABA receptors in the hippocampus which had been associated with the impairment of fear-related memorization facilitate the retrieval of fear-related memory states, and they do this by promoting subcortical activity as opposed to distributed cortical activity typically associated with episodic memory. Of course this makes sense. It serves the survival of an agent to respond reflexively to dangers in the environment, with the somatic marker of fear or rage attaching to those biochemical changes resulting from pre-cortical information processing. This is the pre-reflexive condition into which we awake, and on Fuchs' account this level of embodied memory especially grounds personal identity. Thus, the more or less stable subject over time that, as a relatively regular pattern of activity that

"emerges . . . from a history of embodied experience which has accumulated and sedimented in body memory and as such is implicitly effective in every present moment"⁷⁰ is more or less constituted by unconscious processes, and this has serious implications for any consideration of the "mineness" of consciousness.⁷¹

The ghost is not in the machine. The ghost *is* the machine. Troubles arise when embodied habits do not suit the changing environment, when embodied existence cements history in its "bones" and the traumatized agent can no longer adapt habits to a different habitat. At this extreme, there is no longer any ghost, only mechanism perhaps amounting to a kind of zombie. And here, with the pre-reflective capacity to adjust to environmental changes in the maintenance of prospective integrity, Fuchs points out that these "circular processes" of self regulation are "arguably necessary for the emergence of basic self-awareness" within an artificial consciousness as well. An artificial consciousness must find itself situated in the world, on its way to different situations, with prospective self-awareness and also memory about these situations and the transitions between them.⁷²

With this, Fuchs brings us to a practical limit facing any program in the study of biological consciousness. Every stage of development of an organism embodying Fuchs' deep material memory—and experiencing those peculiarly "mine" moments, for example when a human being completes a paper on consciousness after years of conscious and unconscious preparation—cannot be reliably isolated in the study of a biological system. They may be reported, shared. But, they cannot be exactly reproduced. All of the dimensions weighed in the use of one term rather than another, for example, cannot be systematically tracked. This is not the case for inquiries into artificial conscious systems, however. Indeed, recognizing a similar limit to the biological approach, Edelman, Gally and Baars issue something of a compliment to Boltuc's engineering thesis, advising that we must "accept" that we cannot map cognition in the study of living beings, "to trace causal chains at all levels of complexity in the brain circuits that contribute to consciousness" while at the same time suggesting that a "brain-based device, driven by a simulated brain . . . would be key to success" in understanding consciousness, instead.⁷³

In summary, where Edelman, Gally and Baars recommend research into "brain-based" devices, and Boltuc likewise points to "generators" of consciousness within biological brains, Fuchs suggests that the prospects for artificial consciousness emerge at the interface of embodied cognitive agent and environment, at the level of whole organism in the social historical temporal world that is mine and yours in so far as we embody these horizons. So, to the question "What is a zombie missing?" one might answer "Itself" as a whole.⁷⁴ In the next and final section of this paper we introduce a research program in neurorobotics which instantiates "circular processes" such as those which Fuchs finds requisite for basic self-awareness, leaving the next paper in this series to set out in detail this groups' work in freewill and self-reflective consciousness.

SECTION 4: MINIMAL SELVES

While Boltuc cuts the cognizer into two logical aspects, subject and object, Edelman, Gally and Baars emphasize a dynamic core within a global workspace, and Fuchs finds the subjective and objective standpoints to be together essential to cognition in integrative embodied agency. One thing that all share is a positive assessment of the prospects of a properly configured artificial consciousness, and all generally agree on how such a machine might be built, replicating part or all of an organic system. Following such a recipe in an artificial medium faces difficulties with replicating processing dynamics due to biochemical reality. We will approach these issues in the third paper in our series, when we revisit Boltuc's natural non-reductionism. Current technology does not afford computational power to simulate realistic human brain activity. However, we may not need to instantiate whole brains and narrative-historical political consciousness in chemical metabolisms with all attending systems due natural embodiment in order to isolate aspects of consciousness. Rather, specific features might be drawn in their essential dynamics, such that "a much smaller number of simulated neurons and synapses might prove sufficient to give rise to a particular mental property, such as imagery."⁷⁵

One particularly important aspect of the problem of consciousness as we have drawn it in discussion thus far is the problem of time consciousness, or "temporality", and one especially interesting aspect of temporality is how the raw flow of perceptual experience is parsed and consolidated into narratives composed of sequences of events involving objects as well as other subject agents.⁷⁶ We will describe recent experiments involving the instantiation into robots of this capacity to construct and to deconstruct possible futures, to aim for them so as to explore the consequences in the next paper. Here, by way of introduction, we will briefly review how this research program demonstrates the emergence of "basic self-awareness" in the form of a minimally self-reflective self.

Tani's basic model employs higher and lower levels of differently configured neural networks with the latter tuned to the immediate environment and responsive to rapid changes while the former higher level is attuned to longer ranged patterns. It is in the interactions between these two levels that Tani finds consciousness arising, and he has spent the last two decades building robots which demonstrate this to be the case (for most complete review, see Tani, in press). Here, discussion should turn to the notion of predictive coding and its relationship with the diachronic unity of Tani's neurorobots.⁷⁷

In 1998, Jun Tani detailed a dynamical system structure accounting for the phenomenon of the momentary appearance of the "self" and demonstrated these dynamics in robotics experiments. Tani showed that the self emerges momentarily when the coupled dynamics between the internal neural network and the environment shift from coherent to incoherent dynamics. When everything proceeds as anticipated in the coherent phase, there is no distinction between the self and the environment in the coupled dynamics. However, the self can be perceived as separate from the environment when something goes

wrong, in conflict with the system's anticipation, in the incoherent phase.

In the first experiment (Tani, 1998), constructive and deconstructive interactions between the bottom-up pathway of perception and the top-down pathway of prediction were balanced by internal parameters derived from prior prediction error. Throughout the learning process, the entire system dynamics proceeded with intermittent shifting between the coherent and incoherent phases, with good predictability in the former and poor predictability in the latter. These results were interpreted though Heidegger's (1996) famous example of the hammer, i.e., we become aware of the use of the hammer only when the hammer fails to perform as anticipated, such as when it breaks. In this case, Tani postulated that the gap generated between top-down anticipation and bottom-up reality in the incoherent period represent the difference between the unconscious, routine use of a hammer and its perhaps violent mechanical failure. In this moment, Tani conjectured, the structure of cognitive agency as a "minimal self" rises to awareness. As the agent looks for "What went wrong?" it takes itself as a possible object and answer, "I went wrong." Further, Tani conjectured that the entire system dynamics tends to proceed toward a certain critical state in which a large range of fluctuations may take place, a condition analogous to a system at criticality.⁷⁸

Tani and Nolfi (1999) and Tani (2003) further explored this problem of self-referential selves.⁷⁹ Especially, in a learning experiment with a robot navigating a maze environment (Tani & Nolfi, 1999) and one with a robotic arm manipulating an object (Tani, 2003), the continuous sensorimotor flow of information became segmented into reusable behavior primitives. This chunking was accomplished through a dynamic gate opening/closing (Tani & Nolfi, 1999) or parametric bias shift (Tani, 2003) occurring in a step-wise fashion through the effort of minimizing prediction error, which drove the segmentation or raw perceptual flow into primitive sequences or chunks. After the learning process, the higher level network was also able to predict the sequences of behavior primitives in terms of shifts in this parametric bias vector. Tani interpreted this phenomena as the process of achieving a self-referential self, because the subjective experience of sensorimotor flow becomes objectified into reusable units which are manipulable by higher level processes, e.g., thinking. This interpretation is intuitive, because as the original experience of one's own sensorimotor flow is reconstructed with compositional structures, they become consciously describable objects rather than merely impressions of the original experiences. Then, from this understanding, Tani (2009) found in this capacity the origins of "self-referential selves" as the agent sets out actionable compositions as neurodynamical self-constructs that "emerge ... through self-organizing compositional mechanisms of assembling and de-assembling sensorimotor schemata of repeated experiences", revisionary processes which arise only "in critical conditions of sustaining conflictive interactions between the top-down subjective mind and the bottom-up sensorimotor reality."⁸⁰

The central thesis driving this work has been that consciousness arises in the correction and modification of dynamic structures potentially spontaneously generated in the higher-level cortical area including the pre-frontal cortex (PFC) in biological models. And importantly, a "simulated brain" has not been required to test this thesis. Instead, a much simpler system is able to embody its own possible future situations in the form of a minimal self, and to reconfigure this future as existing projects are frustrated. Tani employs the image of the sandpile, stable with every grain until the last when it all collapses to describe the condition of a dynamic system at a critical point. What results from the sand pile is another sandpile, with potential energy released that was otherwise bound up with its arrangement. The difference between a human being and a sandpile is more or less leisure, metabolism above background, in short a capacity to construct its own order through action. With the complexity of the evolved biological system, we may begin looking at constituent sub-sandpiles and their arrangements, as we had begun with biological systems in the first section of this paper. The questions that remain are merely how many one embodies, in what arrangement, which triggers first and in which contexts.⁸¹

One pressing objection to qualifying any such system as conscious, especially of the hard family of most- and myth-conscious, is that artificial systems are too simple. Of course, simulations of cognition are necessarily less complex than the biological system monitoring them, as overly-complex simulations defeat the purpose of a simulation.⁸² Though it is true that artificial agents in a laboratory *are simpler than* organic brains out functioning in the real world, in artificial consciousness studies the potential exists to isolate essential features with a resolution otherwise lost against the background real-world noise and corruption of the biological approach. These artificial systems may be more conscious, completely conscious, or demonstrate pure consciousness in a way that a biological system cannot, because there are so many facets of cognitive agency essential to living systems that need not be replicated in an artificially conscious system. And, due to the nature of artificial systems, the hurdle that is privileged access to subjectivity may be overcome with perfect information about the dynamic structure of a cognitive system ready at hand. This potential does not exist in the study of biological systems. This potential is afforded, however, by artificial systems as we shall see in greater detail in the next paper in this series.

Artificial conscious systems afford a privileged insight into the structure of cognitive agency and how such systems in their normal operations result in the feelings of being a self in the world, a feeling that meets our every internal self-reflection. These investigations are not limited to available biological models, and there is no risk of polluting the natural environment with genetically engineered creatures designed to represent certain modes of consciousness over others. That said, biological studies continue to inspire artificial systems. For example, Tani's MTRNN architecture⁸³ was inspired by fMRI studies on higher level areas including the PFC showing them important to abstract reasoning and the integration of sensation. Research in mirror neurons

inspired the hypothesis that predictive coding might be essential for pairing generation and recognition of actions, as tested with Tani's RNNPB.⁸⁴ However, the point is that empirical biological results cannot access the core problem of consciousness, as we have tried to articulate in the current paper.

CONCLUSIONS: WHAT TO EXPECT

The next paper in this series details how dynamic complex systems embodied in neurorobots demonstrate consciousness in their normal operations. The third paper in this series will revisit Boltuc's h-, and this paper's most- and myth- consciousnesses in order to evaluate Tani and colleagues' model. Are these robots h-conscious? More? Myth-conscious? At that point, finally we will revisit Boltuc's naturalistic non-reductionist thesis, as it may not be the material nature of a cognitive agent which ultimately grounds any account of consciousness, but rather the dynamic structure that had traditionally only existed in biological, and that now is instantiated also in artificial, forms.

ACKNOWLEDGEMENTS

Special thanks to Peter Boltuc for patient review and deft editing of multiple drafts, as well as to anonymous reviewers of this journal for recommending that we convert one long paper into a series over the course of the year.

NOTES

1. La Mettrie et al., *Man, a Machine: Including Frederick the Great's "Eulogy" on La Mettrie and Extracts from La Mettrie's "The Natural History of the Soul,"* 98–99.
2. Boltuc, "The Philosophical Issue in Machine Consciousness," 159.
3. *Ibid.*, 155.
4. *Ibid.*
5. *Ibid.*, 160.
6. *Ibid.*, 162. In example, he refers to the LIDA model, which posits a cycle of information processing within a Global Workspace, cf. page 157. We will briefly attend to the notion of the "global workspace" in the next section, but generally steer clear of the mire that is competing and complementary accounts of consciousness in the contemporary literature.
7. *Ibid.*, 174.
8. Cf. Boltuc, 160.
9. *Ibid.*, 162.
10. *Ibid.*, 161.
11. *Ibid.*
12. *Ibid.*, 158.
13. *Ibid.*, 162.
14. Cf. Boltuc, 158.
15. *Ibid.*, 173.
16. *Ibid.*, 174.
17. See discussion in *ibid.*, 174.
18. Bitbol, "On the Primary Nature of Consciousness," 268. See also White, *Conscience: Toward the Mechanism of Morality*; "Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience"; "Models of Moral Cognition," which articulate a similar integrative structure in terms of "stitching one's self into the world," and Chalmers, "How Can We Construct a Science of Consciousness?" for extended discussion of problems facing any such account.
19. Boltuc, "The Philosophical Issue in Machine Consciousness," 171.
20. Chalmers, *The Conscious Mind*.
21. The first author is indebted to Alexander VonSchoenborn for educating him to identify these IOUs.
22. Remaining sensitive to the fact that Boltuc's view on this point has developed in the meantime, the current paper attends solely to the position represented in his paper from 2009.
23. Cf. O'Muircheartaigh et al., "White Matter Connectivity of the Thalamus Delineates the Functional Architecture of Competing Thalamocortical Systems."
24. Cf. Llinas et al., "The Neuronal Basis for Consciousness." It is important to note that the systematicity of connected networks is reflected in the neural physiology of non-human animals, as well. See, for example, discussion in O'Muircheartaigh et al., "White Matter Connectivity of the Thalamus Delineates the Functional Architecture of Competing Thalamocortical Systems."
25. Parvizi and Damasio, "Consciousness and the Brainstem," 137.
26. Cf. Boltuc, "The Two Forks in the Road Towards h-consciousness," in press.
27. Bechara, Damasio, and Damasio, "Emotion, Decision Making, and the Orbitofrontal Cortex."
28. Cf. Coppin et al., "When Flexibility Is Stable: Implicit Long-Term Shaping of Olfactory Preferences."
29. Cf. Bechara, Tranel, and Damasio, "Characterization of the Decision-Making Deficit of Patients with Ventromedial Prefrontal Cortex Lesions," 2198.
30. Mitchell et al., "Medial Prefrontal Cortex Predicts Intertemporal Choice," 6.
31. Cf. Janowski, Camerer, and Rangel "Empathic Choice Involves vmPFC Value Signals That Are Modulated by Social Processing Implemented in IPL."
32. Cf. Koenigs et al., "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements."
33. See Mendez, "The Neurobiology of Moral Behavior: Review and Neuropsychiatric Implications," for review.
34. Cf. Damasio, *Self Comes to Mind: Constructing the Conscious Brain*.
35. Cf. Heidegger, *Being and Time: A Translation of Sein und Zeit*, re: "mitda-sein."
36. Cf. Spunt et al., "The Default Mode of Human Brain Function Primes the Intentional Stance."
37. The notion of most-consciousness is much more complicated than Boltuc's original h-consciousness. However, it better reflects the biological reality. Questions remain whether this "thickening" of the analytic sense of h-consciousness is a biological accident or essential to the dynamic structure necessary for cognitive agency in any form. And there is a role for analysis in answering these questions going forward. However, the purpose of analysis is to "carve the world at its joints," rendering complex systems simple enough for ready manipulation, and problems arise when analysis takes inquiry away from the original problem and directs instead to entities that exist only in the context of the analysis. One way to spot these problems is to put the parts together and see if anything is left over or left out. This is the method that we are pursuing in this series of papers, moving from biological to artificial systems integrations. And that said, one dimension of temporal integration remains left out of this model of most-consciousness.
38. Note that I do not write "generates within me the propositional attitude toward" (cf. Thagard, "Desires Are Not Propositional Attitudes").
39. Fuchs ("Self Across Time: The Diachronic Unity of Bodily Existence") recalls these sorts of phenomena as Leibniz's "little perceptions."
40. Note that Boltuc anticipates the form of this discussion, recognizing that there are in any conscious system also "unconscious cognitive processes" on the basis of which "one can ask whether an organism is conscious while still performing

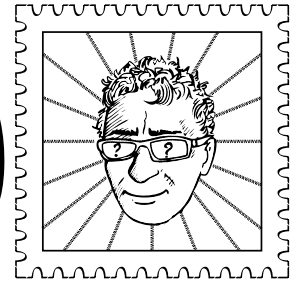
apparent cognitive functions" (Boltuc, "The Philosophical Issue in Machine Consciousness," 160).

41. Again, we must here note that this portrait of h-consciousness is much richer than that of Boltuc's analysis. The view at root here can be found in White, "Models of Moral Cognition."
42. Cf. White, "What Happened? Alcohol, Memory Blackouts, and the Brain," for review.
43. Indeed, the alcohol industry might be characterized as an industry bent on these ends, to some degree, and if zombies were the object, then its potential in this effort is without question.
44. As much as there is a "problem of access" to another consciousness besides one's own, there is the reverse, an obligation to make one's own inner world sensible to others similarly embodied. Where this is impossible, agency is judged differently.
45. As understood, for example, in Habermas, *Communication and the Evolution of Society*.
46. The adolescence of the West after the death of the parents and the mourning period of the mid-life (as if life was a brutal and filthy orphanage), the collective anticipation to get out of that situation arose in the Enlightenment.
47. Boltuc, p. 160
48. Ibid. Note correlate portrait by Watanabe and Mizunami, "Pavlov's Cockroach: Classical Conditioning of Salivation in an Insect."
49. Edelman, Gally, and Baars, "Biology of Consciousness," 5.
50. For instance, implicating the hippocampus in Baars, Franklin, and Ramsoy, "Global Workspace Dynamics: Cortical 'Binding and Propagation' Enables Conscious Contents."
51. Or we may accept that it is only a feature of biological systems.
52. Edelman, Gally, and Baars, "Biology of Consciousness," 2.
53. Fuchs, "Self Across Time."
54. Ibid., 13.
55. Ibid., 8.
56. Ibid., 22.
57. Ibid., 15–16.
58. "We can define the entirety of established habits and skills as implicit or body memory that become current through the medium of the lived body without the need to remember earlier situations"(ibid., 16).
59. Ibid.
60. Cf. Zahavi, "The Experiential Self: Objections and Clarifications."
61. Fuchs, "Self Across Time," 8.
62. Ibid., 22.
63. Edelman, Baars, and Seth, "Identifying Hallmarks of Consciousness in Non-Mammalian Species."
64. It points to something essential so far neglected in our discussion. Take, for example, Aristotle's ideal life of study, ultimately centering on eternal relations and patterns displayed by increasingly perfect beings. Imagine that this is your life and that you wake into its confines. Little is pressing. Life itself is subsumed under an eternal order and your mind works to track this order, rather than remain sensitive to the minutia of daily psycho-political life. Now imagine that you are in an opposite condition, and whatever order on which you had depended in the world has disintegrated in short order. For the cosmological agent whose society may extend to the stars themselves in philosophical reflection on patterns of movement which also extend far beyond the scope of living anticipation, the sense that myth-consciousness extending beyond its embodied constraints (mine, yours, temporal, historical) is the sense into which one would awake. For the terrified agent, there is no anticipation of a future beyond immediate threats and immediately available resources. There is most-consciousness instead.
65. And for an essentially social being with a long early phase of direct mirroring of caregiver action and affect, followed only later with an equally long revisionary period alongside higher level neural growth and entrainment, most habits derive from those witnessed in others more or less alike. There remain questions about how these libraries are collated and how language emerges in distinct ways, permitting resonance between the similarly embodied while remaining exclusive of others. These questions can be answered within the context of artificial consciousness studies while remaining practically outside of any biological inquiry. For organisms like human beings, experience is historical-material of necessity. Where this is missing, there is deficiency. And due to deep material memory furnishing future self-situations extending well beyond any that an organism may itself potentially inhabit, for example, a just world after a few generations of adjustment for a childless man today, each moment is located within variable horizons of anticipation of succeeding moments against retained past, and not only for one's self and those alike, but potentially for any consciousness at all.
66. Fuchs, "Self Across Time," 12.
67. Ibid., 16.
68. Ibid., 17; quoting from Applefield, *The Story of a Life: A Memoir*, 90.
69. Jovasevic et al., "GABAergic Mechanisms Regulated by miR-33 Encode State-Dependent Fear."
70. Fuchs, "Self Across Time," 18.
71. See Abitbol et al., "Neural Mechanisms Underlying Contextual Dependency of Subjective Values: Converging Evidence from Monkeys and Humans," for review of common biological mechanisms.
72. Fuchs expresses doubts that "processes of vital self-regulation in the brainstem and diencephalon" may be replicated in an artificial agent, not solely due to their own complexity but because the brain is embedded in "rather slow biochemical interactions" which "may not be described as digital information processing in analogy to a computer" (14). In an artificial agent, these processes must take on a different form, and to which degree they need to be replicated remains to be seen.
73. Edelman, Gally, and Baars, "Biology of Consciousness," 5.
74. Implications of differing material modes of embodiment as this self is constituted are weighed in the third paper of this series.
75. Edelman, Gally, and Baars, "Biology of Consciousness," 5.
76. Cf. Tani and Nolfi, "Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems"; Tani, "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study"; and Tani, *Exploring robotic minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*
77. We set aside these issues for the next paper in order to quickly describe how a minimal self arises as a result of the basic interactions between these two embodied temporalities, and thus how any such system may appear as a "diachronic unity" in the first place.
78. Bak et al., "Self-Organized Criticality: An Explanation of the 1/ fnoise."
79. Tani and Nolfi, "Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems"; Tani, "Learning to Generate Articulated Behavior Through the Bottom-Up and the Top-Down Interaction Process."
80. Tani, "Autonomy of 'Self' at Criticality: The Perspective from Synthetic Neuro-Robotics," 423.
81. What is it like to be a sandpile.
82. Cf. White, "Simulation, Self-Extinction, and Philosophy in the Service of Science."
83. Yamashita and Tani, "Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: A Humanoid Robot Experiment."
84. Tani, "Learning to Generate Articulated Behavior Through the Bottom-Up and the Top-Down Interaction Process."

REFERENCES

- Abitbol, R., M. Lebreton, G. Hollard, B. J. Richmond, S. Bouret, and M. Pessiglione. "Neural Mechanisms Underlying Contextual Dependency of Subjective Values: Converging Evidence from Monkeys and Humans." *Journal of Neuroscience* 35, no. 5 (2015): 2308–20.
- Appelfeld, A. *The Story of a Life: A Memoir*. New York: Random House, 2009.
- Bak, P., C. Tang, and K. Wiesenfeld. "Self-Organized Criticality: An Explanation of the 1/fnoise." *Physical Review Letters* 59, no. 4 (1987): 381–84.
- Baars, B. J. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, 1997.
- Baars, B. J., S. Franklin, and T. Z. Ramsay. "Global Workspace Dynamics: Cortical 'Binding and Propagation' Enables Conscious Contents." *Frontiers in Psychology* 4 (2013): 1–22.
- Bechara, A., H. Damasio, and A. R. Damasio. "Emotion, Decision Making, and the Orbitofrontal Cortex." *Cerebral Cortex* 10, no. 3 (2000): 295–307.
- Bechara, A., D. Tranel, and H. Damasio. "Characterization of the Decision-Making Deficit of Patients with Ventromedial Prefrontal Cortex Lesions." *Brain: A Journal of Neurology* 123 (2000): 2189–202.
- Bitbol, M. "On the Primary Nature of Consciousness." In *The Systems View of Life*, edited by F. Capra and P. L. Luisi, 266–68. Cambridge: Cambridge University Press, 2014.
- Boltuc, P. "The Philosophical Issue in Machine Consciousness." *International Journal of Machine Consciousness* 1, no. 1 (2009): 155–76.
- Boltuc, P. "The Two Forks in the Road Towards h-consciousness." *Theoria et Historia Scientiarum* (in press, 2016).
- Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press, 1996.
- Chalmers, D. J. "How Can We Construct a Science of Consciousness?" *Annals of the New York Academy of Sciences* 1303, no. 1 (2013): 25–35.
- Coppin, G., S. Delplanque, C. Porcherot, I. Cayeux, and D. Sander. "When Flexibility Is Stable: Implicit Long-Term Shaping of Olfactory Preferences." *PLoS ONE* 7, no. 6 (2012): e37857. <http://doi.org/10.1371/journal.pone.0037857>
- Damasio, A. R. *Self Comes to Mind: Constructing the Conscious Brain*. New York: Pantheon Books, 2010.
- Edelman, D. B., B. J. Baars, and A. K. Seth. "Identifying Hallmarks of Consciousness in Non-Mammalian Species." *Consciousness and Cognition* 14, no. 1 (2005): 169–87.
- Edelman, G. M., J. A. Gally, and B. J. Baars. "Biology of Consciousness." *Frontiers in Psychology* 2, no. 4 (2011): 1–7. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3111444>
- Fuchs, T. "Self Across Time: The Diachronic Unity of Bodily Existence." *Phenomenology and the Cognitive Sciences* (forthcoming, 2016) doi: 10.1007/s11097-015-9449-4
- Habermas, J. *Communication and the Evolution of Society*. Boston: Beacon Press, 1979.
- Heidegger, M., and J. Stambaugh. *Being and Time: A Translation of Sein und Zeit*. Albany, NY: State University of New York Press, 1996.
- Janowski, V., C. Camerer, and A. Rangel. "Empathic Choice Involves vmPFC Value Signals That Are Modulated by Social Processing Implemented in IPL." *Social Cognitive and Affective Neuroscience* 8, no. 2 (2013): 201–08.
- Jovasevic, V., K. A. Corcoran, K. Leaderbrand, N. Yamawaki, A. L. Guedea, H. J. Chen, G. M. G. Shepherd, and J. Radulovic. "GABAergic Mechanisms Regulated by miR-33 Encode State-Dependent Fear." *Nature Neuroscience* 18, no. 9 (2015): 1265–71.
- Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, and A. Damasio. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements." *Nature* 446, no. 7138 (2007): 908–11.
- La Mettrie, J. O., G. C. Bussey, M. W. Calkins, and Frederick, King of Prussia. *Man, a Machine: Including Frederick the Great's "Eulogy" on La Mettrie and Extracts from La Mettrie's "The Natural History of the Soul."* Chicago, Illinois: Open Court Publishing Co., 1912.
- Llinas, R., U. Ribary, D. Contreras, and C. Pedroarena. "The Neuronal Basis for Consciousness." *Philosophical Transactions: Biological Sciences* 353, no. 1377 (1998): 1841–49.
- Mendez, M. F. "The Neurobiology of Moral Behavior: Review and Neuropsychiatric Implications." *CNS Spectrums* 14, no. 11 (2009): 608–20. Last accessed January 31, 2016, at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163302/>
- Mitchell, J. P., J. Schirmer, D. L. Ames, and D. T. Gilbert. "Medial Prefrontal Cortex Predicts Intertemporal Choice." *Journal of Cognitive Neuroscience* 23, no. 4 (2011): 857–66.
- O'Muircheartaigh, J., S. S. Keller, G. J. Barker, and M. P. Richardson. M. P. "White Matter Connectivity of the Thalamus Delineates the Functional Architecture of Competing Thalamocortical Systems." *Cerebral Cortex* 25, no. 11 (2015): 4477–89.
- Parvizi, J., and A. Damasio. "Consciousness and the Brainstem." *Cognition* 79 (2001): 135–59.
- Spunt, R. P., M. L. Meyer, and M. D. Lieberman. "The Default Mode of Human Brain Function Primes the Intentional Stance." *Journal of Cognitive Neuroscience* 27, no. 6 (2015): 1116–24.
- Tani, J. "An Interpretation of the 'Self' from the Dynamical Systems Perspective: A Constructivist Approach." *Journal of Consciousness Studies* 5, nos. 5-6 (1998): 516–42.
- Tani, J., and S. Nolfi. "Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems." *Neural Networks*, 12, no. 7 (1999): 1131–41.
- Tani, J. "Learning to Generate Articulated Behavior Through the Bottom-Up and the Top-Down Interaction Process." *Neural Networks* 16 (2003): 11–23.
- Tani, J., and M. Ito. "Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics: A Robot Experiment." *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* 33, no. 4 (2003): 481–88.
- Tani, J. "The Dynamical Systems Accounts for Phenomenology of Immanent Time: An Interpretation by Revisiting a Robotics Synthetic Study." *Journal of Consciousness Studies* 11, no. 9 (2004): 5–24.
- Tani, J. "Autonomy of 'Self' at Criticality: The Perspective from Synthetic Neuro-Robotics." *Adaptive Behavior* 17, no. 5 (2009): 421–43.
- Tani, J. *Exploring robotic minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. New York: Oxford University Press, forthcoming.
- Thagard, P. "Desires Are Not Propositional Attitudes." *Dialogue: Canadian Philosophical Review / Revue Canadienne De Philosophie* 45, no. 1 (2006): 151–56.
- Watanabe, H., and M. Mizunami. "Pavlov's Cockroach: Classical Conditioning of Salivation in an Insect." *PLoS ONE*, 2, no. 6 (2007): e529. doi:10.1371/journal.pone.0000529
- White, A. M. "What Happened? Alcohol, Memory Blackouts, and the Brain." *Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism* 27, no. 2 (2003): 186–96.
- White, J. B. *Conscience: Toward the Mechanism of Morality*. Columbia, MO: University of Missouri–Columbia, 2006.
- White, J. "Understanding and Augmenting Human Morality: An Introduction to the ACTWith Model of Conscience." *Studies in Computational Intelligence* 314 (2010): 607–21.
- White, J. "Models of Moral Cognition." In *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, edited by L. Magnani, 363–91. Berlin: Springer, 2014.
- White, J. "Simulation, Self-Extinction, and Philosophy in the Service of Science." *AI & Society* (2015) doi:10.1007/s00146-015-0620-9
- Yamashita, Y., and J. Tani. "Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: A Humanoid Robot Experiment." *PLoS Computational Biology* 4, no. 11 (2008): e1000220.
- Zahavi, D. "The Experiential Self: Objections and Clarifications." In *Self, No Self?: Perspectives from Analytical, Phenomenological, and Indian Traditions*, edited by M. Siderits, E. Thompson, and D. Zahavi. Oxford: Oxford University Press, 2010.

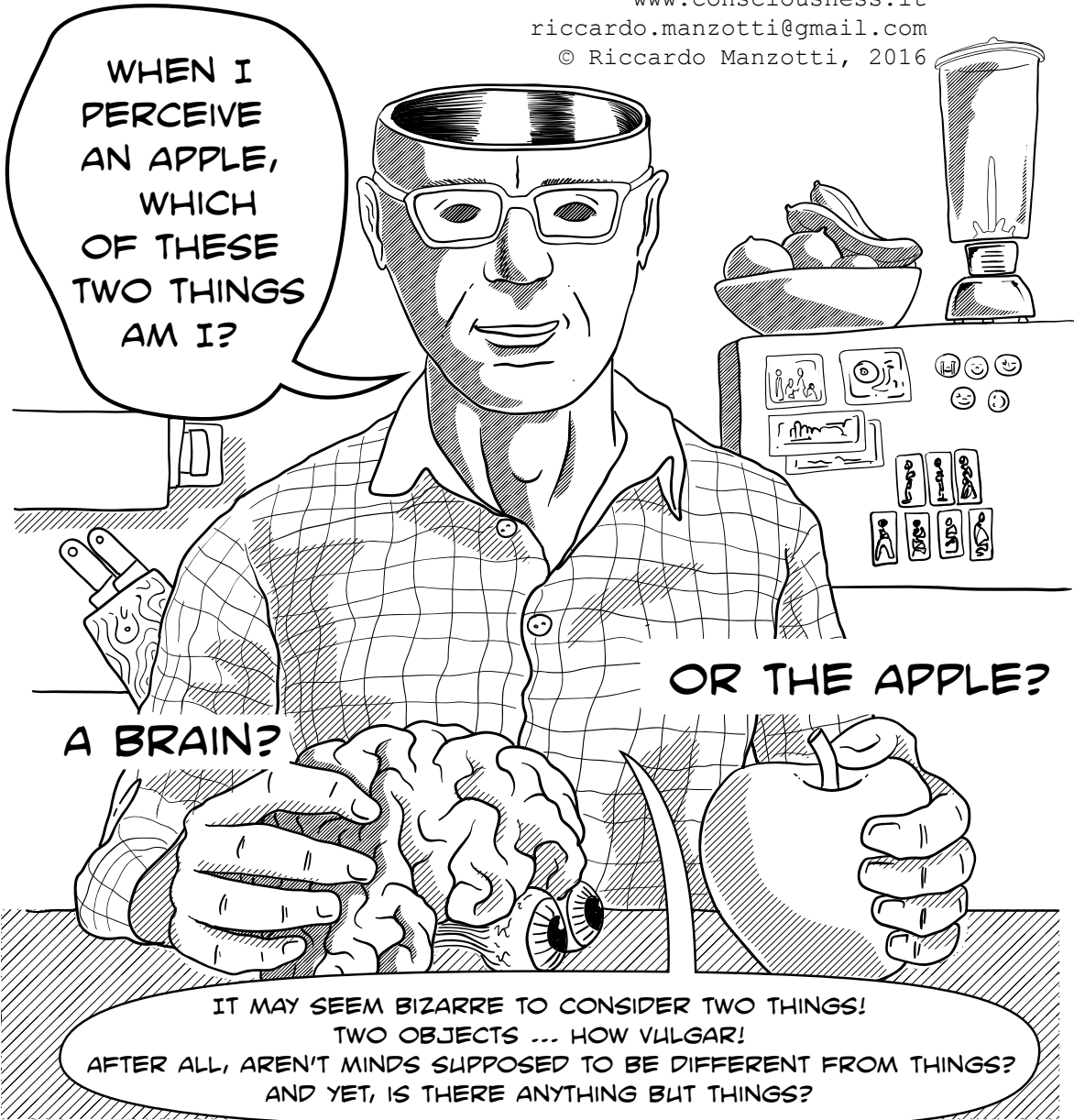
OFF BOUND

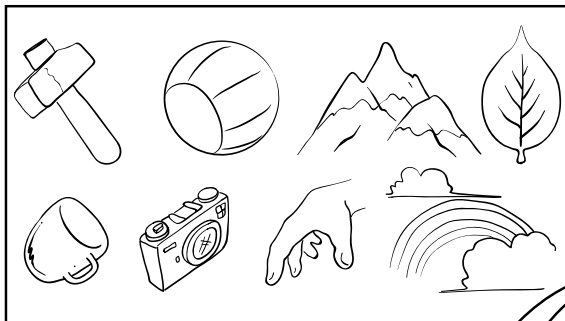


A MIND OBJECT

IDENTITY THEORY

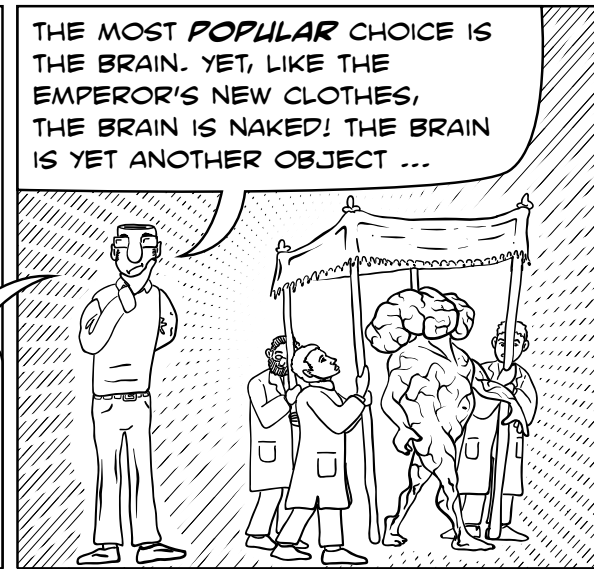
www.consciousness.it
riccardo.manzotti@gmail.com
© Riccardo Manzotti, 2016



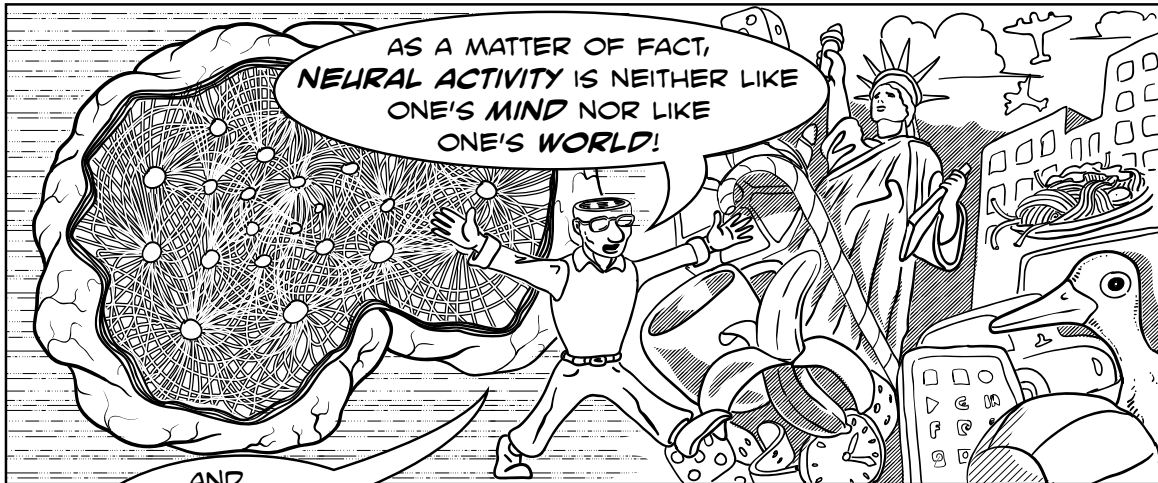


THE MOST **POPULAR** CHOICE IS THE BRAIN. YET, LIKE THE EMPEROR'S NEW CLOTHES, THE BRAIN IS NAKED! THE BRAIN IS YET ANOTHER OBJECT ...

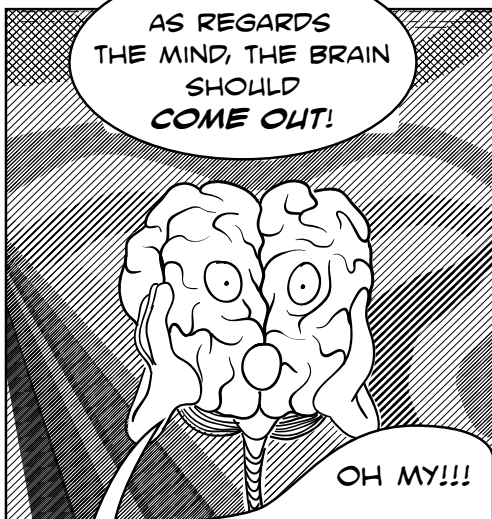
I MEAN 'THINGS' BECAUSE IN NATURE EVERYTHING IS A **SPATIOTEMPORALLY** ENTITY MADE OF **MATTER/ENERGY** WITH A **CAUSAL** ROLE, THAT IS ... A **THING!**



AS A MATTER OF FACT, **NEURAL ACTIVITY** IS NEITHER LIKE ONE'S **MIND** NOR LIKE ONE'S **WORLD!**



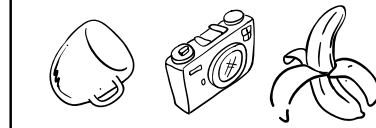
AND ... AS REGARDS THE MIND, THE BRAIN SHOULD **COME OUT!**



OH MY!!!


I AM A **PHYSICAL** OBJECT I THOUGHT I WAS **SPECIAL**, BUT I AM NOT!

FURTHERMORE, IF OBJECTS HAD NONE OF EXPERIENCE'S PROPERTIES



ARE **PHYSICAL**

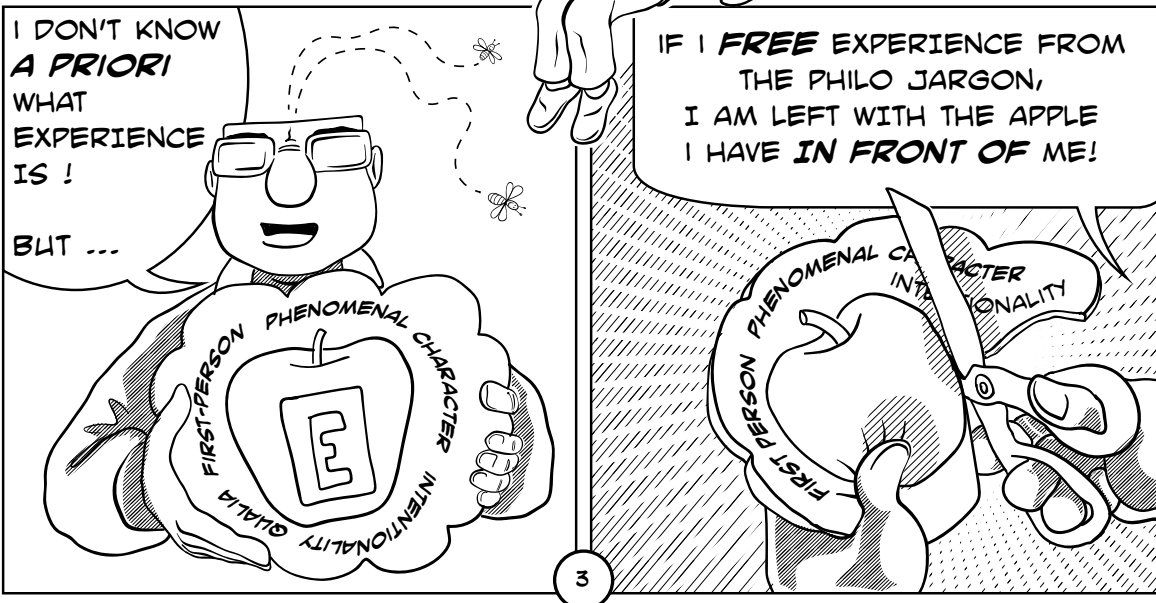
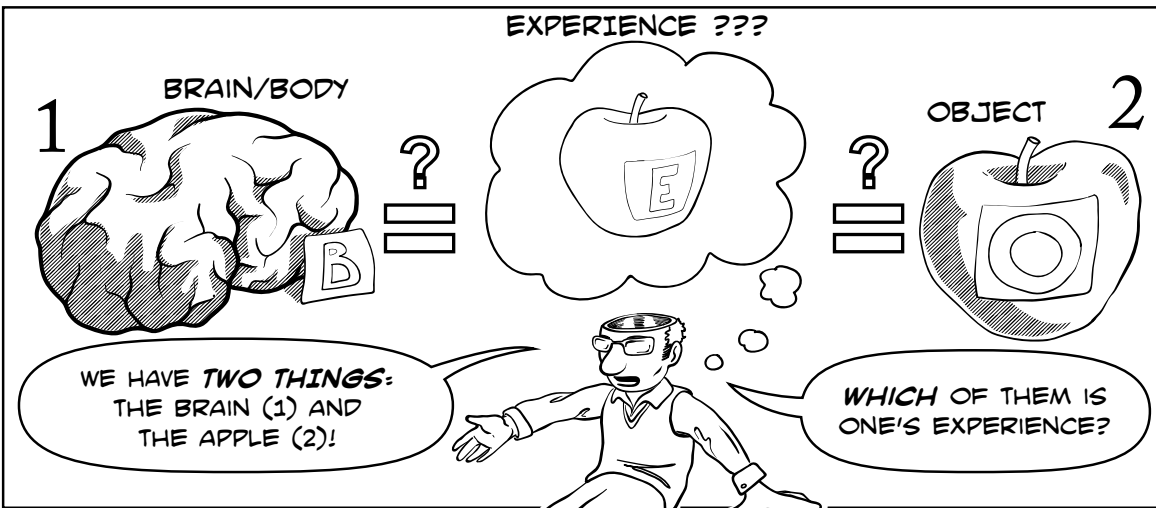
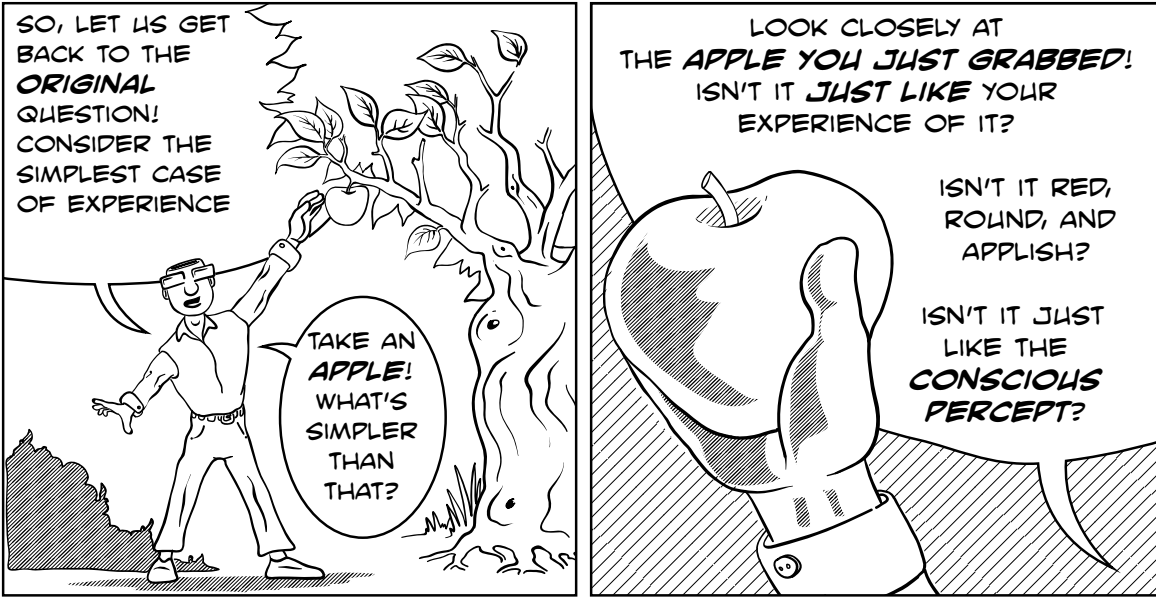
THEN, THE BRAIN, BEING AN OBJECT, COULD HAVE NONE EITHER!



IS **PHYSICAL** TOO!

HOWEVER, RATHER THAN THROW AWAY THE BRAIN, LET'S CONSIDER

2 OTHER **PHYSICAL** CANDIDATES!



THUS, IF I CONTRAST THE **EXPERIENCE** OF THE APPLE WITH THE **APPLE**, THEY ARE THE **SAME!**

ON THE OTHER HAND, **BRAIN** AND THE **EXPERIENCE** OF THE APPLE ARE COMPLETELY **DIFFERENT!**

I CAN **PUT BACK** MY BRAIN IN MY SKULL, THE BRAIN IS THE ONLY OBJECT I AM NOT! I MUST BE ELSEWHERE!

ONE MIGHT PUT EXPERIENCE **OUTSIDE** OF THE PHYSICAL WORLD. BUT IT WOULD BE A **NON STARTER**

HERE?
IN ANOTHER DIMENSION?

SOLUTION 1: DUALISM

WHERE AM I THEN? WHERE IS THE EXPERIENCE OF THE APPLE?

OR, ONE MIGHT PUT THE MIND INSIDE THE BRAIN. BUT, THE BRAIN IS JUST AN OBJECT AND INSIDE THE HEAD **NOTHING** IS LIKE ONE'S EXPERIENCE

HERE?

THEN, I AM LEFT WITH THE ONLY PHYSICAL CANDIDATE: THE OBJECT! A **MIND-OBJECT** IDENTITY THEORY

OR, SURPRISINGLY, HERE?

IN SHORT THE IDEA IS THAT ONE'S EXPERIENCE IS THE **VERY** OBJECT ONE EXPERIENCES! CALL IT **OBJECT BOUND**

I AM THE MIND!!!

OBJECT = EXPERIENCE

SOLUTION 2: BRAINBOUND

SOLUTION 3: OBJECTBOUND!

4

LET'S TAKE A QUANTUM LEAP! LEAVE BEHIND THE ORTHODOX BRAIN/BODY-CENTRIC VIEW OF REALITY AND EMBRACE A NEW STANCE. THE CENTER IS NO LONGER THE BODY BUT THE OBJECT!

THE OLD VIEW!

'PTOLEMAIC' BRAIN-CENTERED VIEW OF THE MIND

THE HYPOTHESIS IS RATHER STRAIGHTFORWARD!

WHEN I EXPERIENCE AN OBJECT, *THE THING I AM* IS THE VERY OBJECT! IT IS A *MIND-OBJECT IDENTITY THEORY!*

THE NEW VIEW!

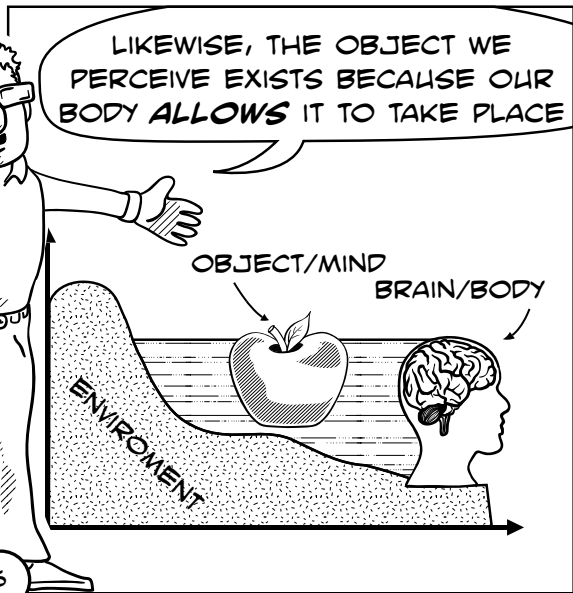
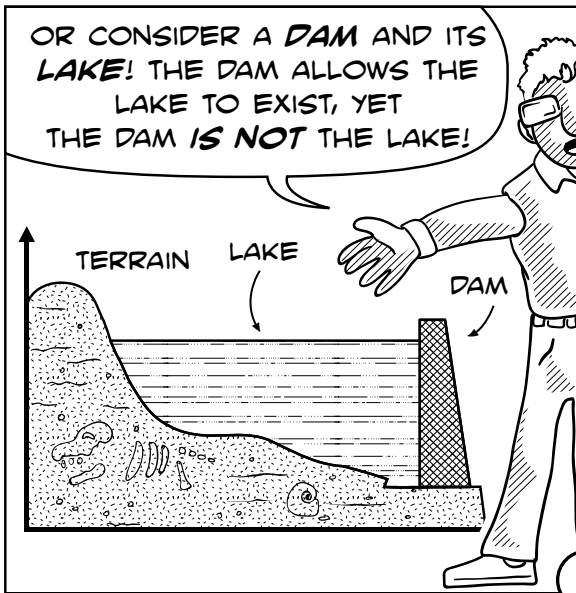
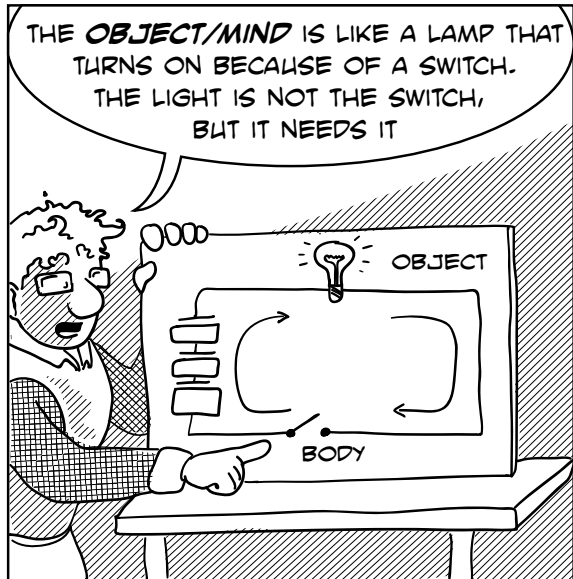
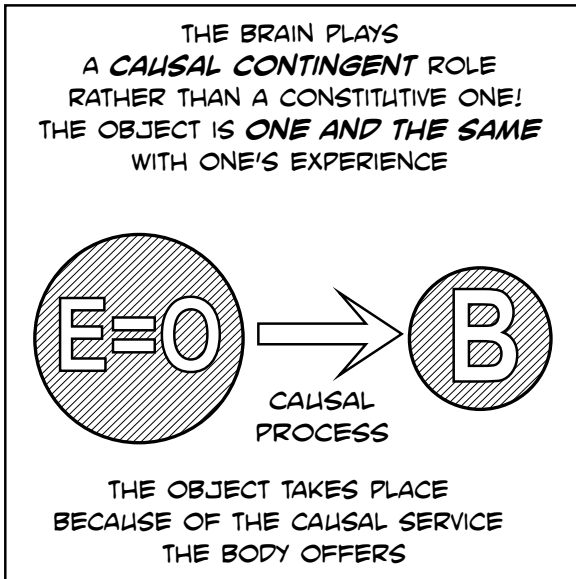
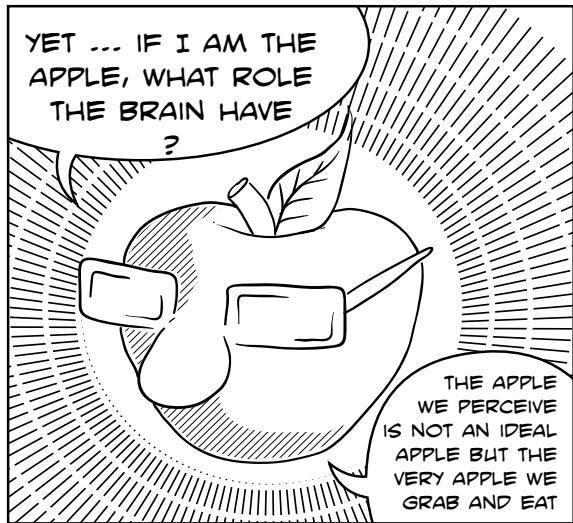
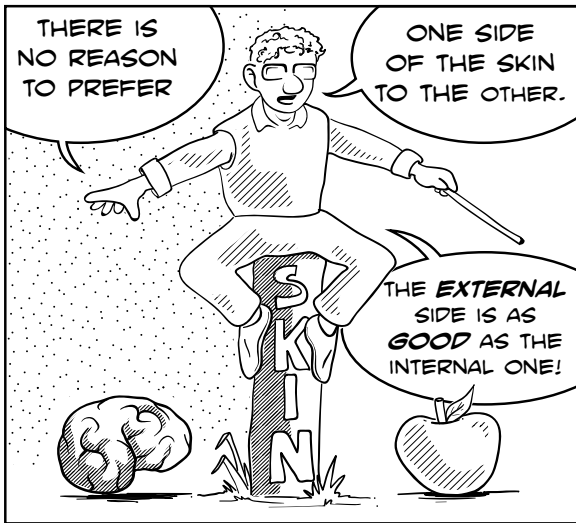
'COPERNICAN' OBJECT-CENTERED VIEW OF THE MIND

WAIT A SEC!
I CAN'T BE THE APPLE!
I FEEL I AM HERE! INSIDE MY HEAD! **BEHIND** MY EYES AND **BETWEEN** MY EARS!

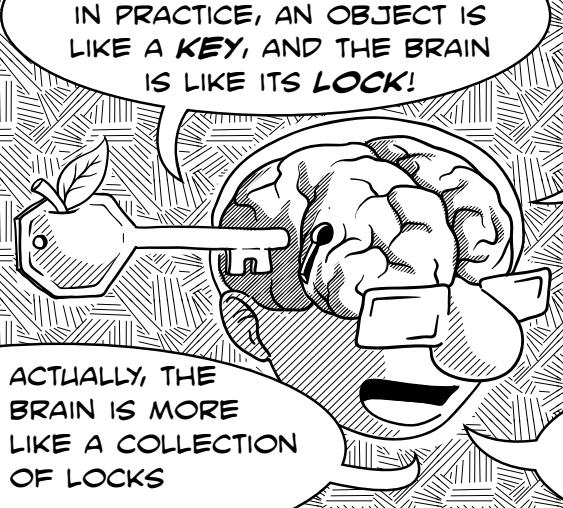
MY MIND IS NOT HERE

NONSENSE!
ONE DOES NOT FEEL *WHERE* THE MIND IS, ONE FEELS *WHERE THE BODY IS!*
MORE PRECISELY, WHERE SENSORY ORGANS ARE.

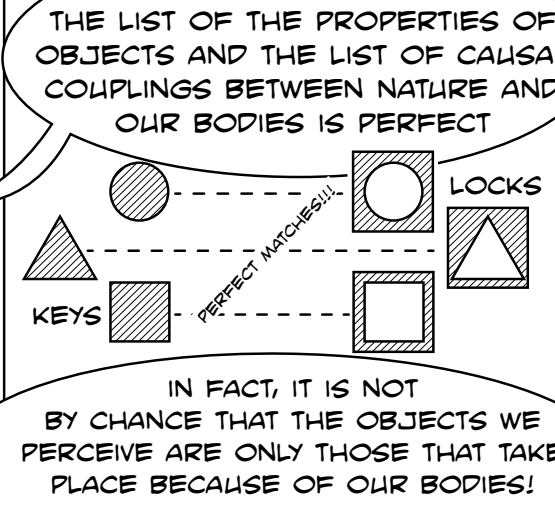
DENNETT SHOWED HOW TO DEBUNK SUCH A NOTION!



IN PRACTICE, AN OBJECT IS LIKE A **KEY**, AND THE BRAIN IS LIKE ITS **LOCK**!



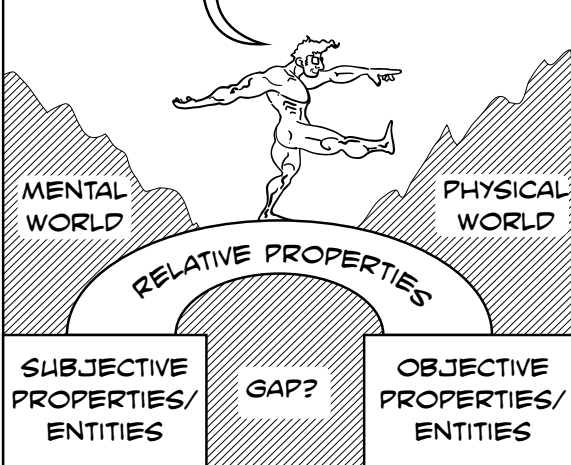
THE LIST OF THE PROPERTIES OF OBJECTS AND THE LIST OF CAUSAL COUPLINGS BETWEEN NATURE AND OUR BODIES IS PERFECT



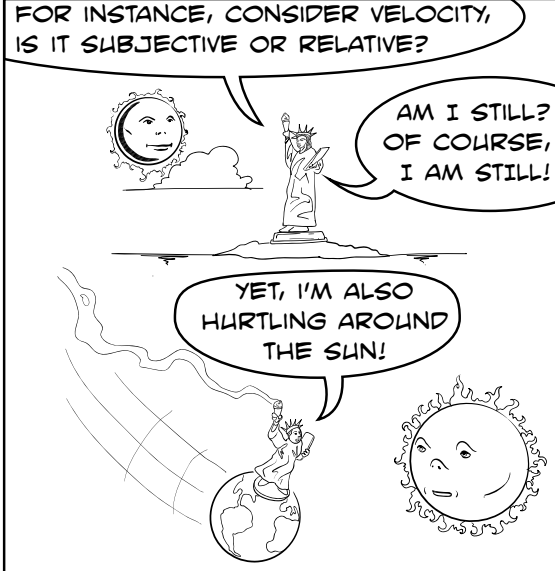
ACTUALLY, THE BRAIN IS MORE LIKE A COLLECTION OF LOCKS

IN FACT, IT IS NOT BY CHANCE THAT THE OBJECTS WE PERCEIVE ARE ONLY THOSE THAT TAKE PLACE BECAUSE OF OUR BODIES!

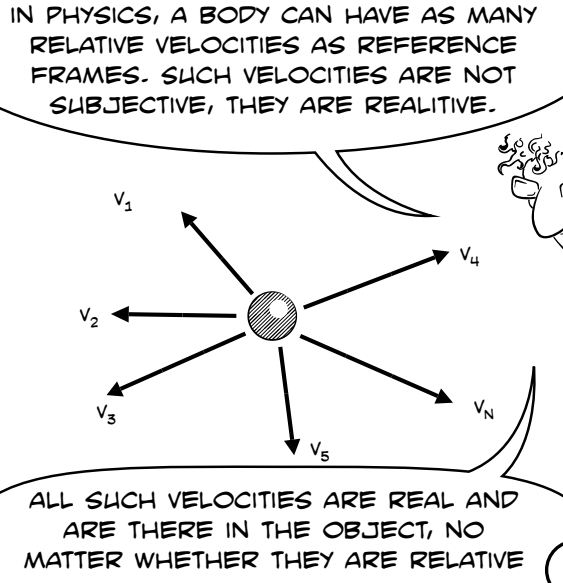
THUS, WE CAN FINALLY OVERCOME THE DREADED SUBJECTIVE VS. OBJECTIVE GAP (AKA THE HARD PROBLEM)!



FOR INSTANCE, CONSIDER VELOCITY, IS IT SUBJECTIVE OR RELATIVE?

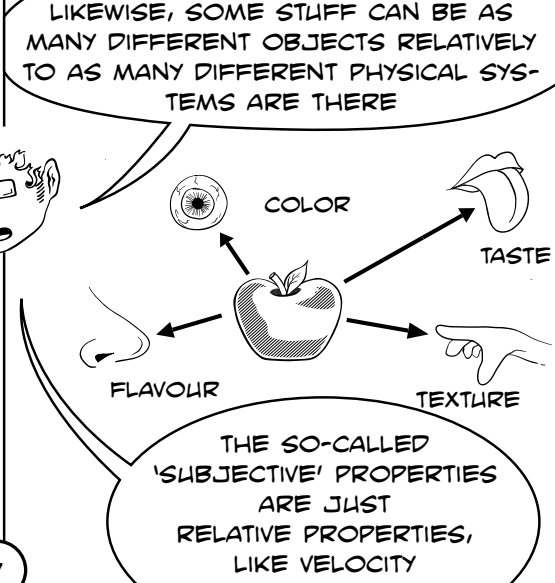


IN PHYSICS, A BODY CAN HAVE AS MANY RELATIVE VELOCITIES AS REFERENCE FRAMES. SUCH VELOCITIES ARE NOT SUBJECTIVE, THEY ARE RELATIVE.



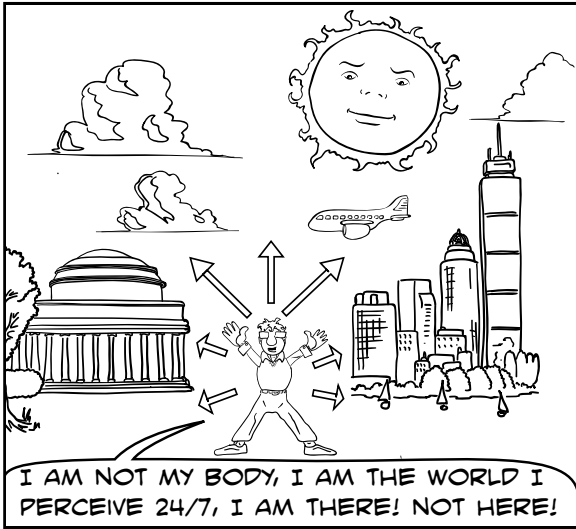
ALL SUCH VELOCITIES ARE REAL AND ARE THERE IN THE OBJECT, NO MATTER WHETHER THEY ARE RELATIVE

LIKewise, SOME STUFF CAN BE AS MANY DIFFERENT OBJECTS RELATIVELY TO AS MANY DIFFERENT PHYSICAL SYSTEMS ARE THERE

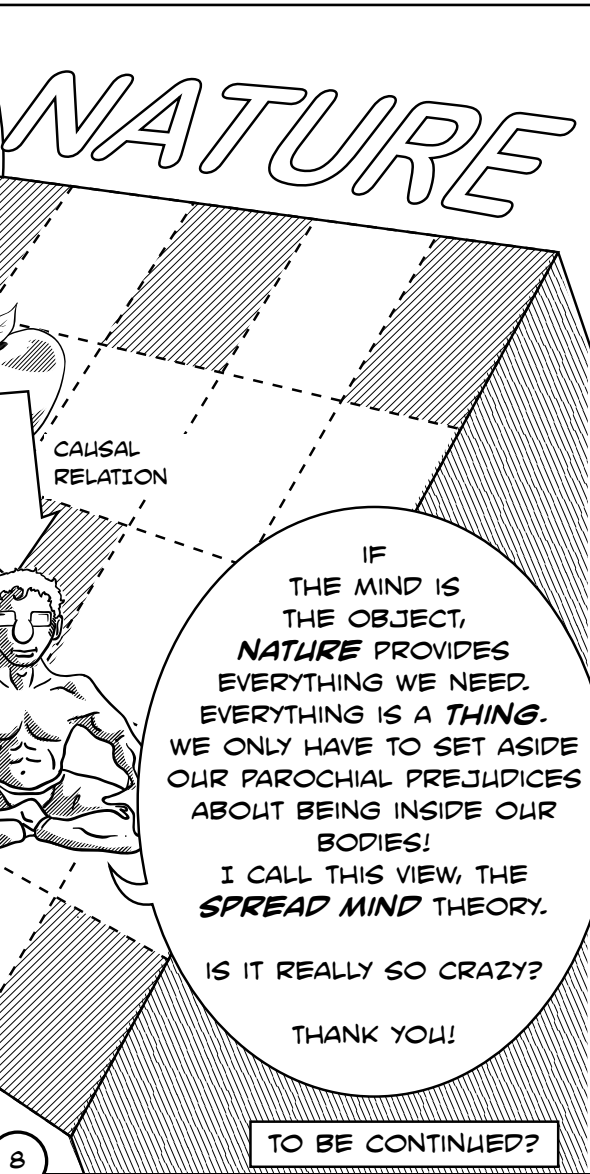


THE SO-CALLED 'SUBJECTIVE' PROPERTIES ARE JUST RELATIVE PROPERTIES, LIKE VELOCITY

7



NATURE IS A SPATIOTEMPORAL MANIFOLD POPULATED OF **THINGS** WITH A CAUSAL ROLE: THE BODY AND THE OBJECT. WHICH ONE IS MY EXPERIENCE? THE IDEA IS REALLY **SIMPLE**: THE EXPERIENCE OF THE APPLE IS THE APPLE! I AM THE OBJECT. MY MIND IS WHERE THE OBJECT IS! AND NOT WHERE THE BODY IS. THE BODY AND THE MIND ARE IN TWO **DIFFERENT** PLACES.



© Riccardo Manzotti, 2016
 IULM University, Milan
 www.thespreadmind.com
 www.consciousness.it
 riccardo.manzotti@gmail.com

On Importance of Life and Death for Artificial Intelligent Creatures

Jordi Vallverdú

UNIVERSITY AUTONOMOUS OF BARCELONA

Max Talanov

KAZAN FEDERAL UNIVERSITY

INTRODUCTION

One of the moments that defines the life of any living system, with some higher cognitive skills, is the crucial instant during which that being becomes aware of its own death. There is a corollary of this idea: the inevitable and large physical decrepitude of the body (and mind) that usually is related to aging processes. Our biological clocks, though we could ignore them for some time, never stop counting our minutes of the active life. Despite the fact that people know about their own death, even at early child stages¹ only sudden, accidental, traumatic and unexpected signals of this fact produce changes into the daily lives of those humans.² Think, for example, of how Siddharta Gautama Buddha reacted to the awareness of the existence of death after being deprived of this information by his overprotective father. The typical midlife crisis of the '40s and '50s is another example of this phenomenon. Psychologists have studied how humans tend to show attitudes towards death, trying to analyze and even quantify it, producing measurements like Death Anxiety Scale (DAS),¹ Fear of the Death Scale (FDS),² or Death Concern Scale (DC).³ And especially in those cases of survival of unexpected death peril (human menace, illness, natural disaster, etc.), humans tend to completely change their values and lives. Healing from trauma is commonly defined by those who have experienced it as a "rebirth" process. The life on earth and development of the societies—including literature, arts, and all the political apparatus—would have appeared dramatically different in absence of aging and death. It is therefore legitimate, in times of optimism regarding AI development, to consider how influential aging and death (in the sense of awareness of it) could be for the development of a future hybrid society where machine and biological intelligence can productively cooperate. Taking into account the cognitive mechanisms related to aging and death, we ask ourselves how a new generation of realistic Artificial Intelligence Agents (henceforth, AIA) could cope with their own degeneration, malfunction, and even final disintegration or death. In spite of of radical life extension programs and expectations, the truth is that inside of an entropic universe, the final and big death (the Big Freeze, Big Rip, Re-explosion, as possible hypotheses) are inevitable.

MAPPING THE EXISTENTIAL CHALLENGES OF ARTIFICIAL ADAPTIVE INTELLIGENT SYSTEMS

We review several scenarios that an AIA can find once monitoring its own embodiment as well as the environment in which it will need to operate. This way we could be able to design future mechanisms to control, nurture, and train these

machines correctly as well as to prepare for their possible reactions. For example, the so-called more intelligent AI program nowadays, ConceptNet system, has a higher IQ than a four-year-old, and it can face this kind of problem. It will also help us to improve the rational approach to AI performance and actions.⁶ Consequently, death awareness and aging are the two related phenomena that we analyze in this paper. Our idea is that in spite of the fact of possible repairs, definitive extension and maintenance of dynamic cognitive structures are not possible. Aging and death seem to be necessary for conscious beings. Besides, constant learning rates and open attitudes towards new ideas are not possible for the systems with strong beliefs and experiential results. Any homeostatic system tends to equilibrate the informational universe under its range of possible decisions and actions.⁷ Figure 1 shows qualitatively the development and decay of cognitive functions via maturation and senescence, described in detail below.

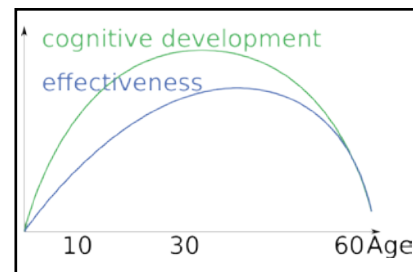


Figure 1. The cognitive development and decay.

Consequently, we can define the following scenarios in which an AIA should be faced to aging/death problems:

1. *Mechanical*: bodies get older, and their parts become worn off. Some can be changed, but some others not. Brains, for example, are not transferable and even lesser minds lose their properties when transferred. Professor Warwick created a robot that experienced problems with biological brains implemented into robots: these brains need to adapt to the new embodiments, but at the same time, the brains became older and needed to be replaced.⁸ The mixture of living and artificial mechanisms is something that is faced with serious problems from the eternal life perspective. Even in the case of considering classic robots, due to their hundreds or thousands of components, in case of self-development and self-adaptation similar to a human brain, the easy transfers will not be possible once their compounds experience life-cycle influence and aging and malfunction . . . like the beings do.
2. *Informational*: At the same time, we infer that these intelligent creatures will be faced with some important formal-architectural problems: first of all, the current Von Neumann architecture is inadequate for the reimplementation of neuronal structures in a computational machine. Most promising simulations are now done in realistic neural networks like NEST. The inability to simulate

the neuronal process in real time in realistic scales of mammalian brain triggers the interest of scientists towards non-Von Neumann architectures, possibly the brightest example of which is the SyNAPSe project of IBM, which recently provided the novel chip technology TrueNorth. It was declared that this kind of technology could provide the proper electronic basement for self-learning approaches implemented in the hardware directly. This is an example of approaches towards electronic minds comparable with human. We have to mention possible economic problems involved into these minds: patents on algorithmic approaches.⁹

3. *Cognitive*: any current cognitive system which is able to learn and to be creative follows some operational architectures that at any extent reproduce human cognitive skills. Some extraordinary expert systems, from classic Logic Theorist to contemporary Deep Blue or Watson (both by IBM), have achieved great results in specialized domains. But the distance between expert systems following formal rules and real humans is quite big. Human beings are epistemic opportunists¹⁰ and this is at the same time the result of conscious strategies as well as of internal cognitive mechanisms that we could call as "bounded."¹¹ In order to take a decision, we follow several steps, strongly influenced by environmental, personal, or cultural factors. This could be understood as the basis for the explanation that humans feel tired, bored (creativity requires curiosity, revised goals, etc.), procrastination, laziness, depression (and possible suicide), and a long list of attitudinal dispositions. Thus, a creative, innovative and audacious machine should be able to create its own reasoning rules and strategies, being at the same time influenced by several informational conditions (internal and external).
4. *Epistemic*: intelligent machines will be faced with some limits of our informational sphere: quantum underdetermination, multilayered or systemic approach to any event of the world, formal limitations (think, for example, of Russell's paradox or Gödel's incompleteness theorems), lack of data, impossibility of computing all possible outcomes of some events. This could force this kind of AI systems to a dead-end of the epistemic decisions: whether the activity is meaningful or not. For example, the belief in existence of nature as a real thing made the birth of empirical sciences possible. In contrast, intelligent people who considered nature as the result of our "mistaken senses" also decided to embrace a more religious or supranaturalistic approach to nature. Thus, we can conclude that the epistemic horizon determines the organization and evaluation of information-seeking activities. In that sense, these epistemological views reinforce or generate special ontological approaches to the reality. It creates a vicious circle that increases these bonds each year. We should assume that the older AI

should not use the revolutionary steps to change, tending to a stabilization and lack of progress or paradigmatic change. Scientific revolutions are held by young or middle-age agents who try to find their place in academia and society. Therefore, the different kind of intelligent activities (which include linguistic tools, action organization, or epistemic methods, among others), always require some kind of social cooperation.¹²

5. *Emotional*: moods and emotional clues: a long-living AI should have more time to think, but intrinsic limitations of our universe can depress it: Big Freeze, loss of energy by entropy, stars consumed, galaxies' constant expansion until the context of an universe absolutely black (no star lights able to cross the whole universe distance, so no possibility of reaching any new place). It would imply the certainty of a death eventually and possibly lack of sense, modifying their mood and behavior. At the same time, we need to consider the instabilities that such emotional architectures would imply for the behavior and goal definitions for older AI systems.
6. *Moral*: we are now facing that an AIA will need to deal directly with moral decisions over death and life: intelligent cars, medicine decisions, military robots, etc. This has a strong relationship with the fact: elderly populations tend to be more conservative in their decisions.¹³ For the same opposite cognitive (and hormonal) reasons, young people are impetuous, brave, prone to risky or unknown activities. This is not only related to biochemical aspects of the body, but also to the knowledge acquisition as well as the necessity of generating stable conceptual frameworks and optimizing life expectancy. Consequently, these aged AIA would tend to defend conservative moral aspects instead of other, much more dynamic, ones. This could affect the machine's decisions about medical care, human surveillance, combat strategies, or industrial organization, among others.
7. *Consciousness*: the Japanese shintoist idea, "mono no aware," that can be translated as "an empathy toward things, or "a sensitivity to ephemera." It constitutes the basis of haiku poetry, which is based on the impression into some human soul of a special moment, which is not necessarily so special for any other human being. It allows the generation of poems, because they try to connect special experience with the description of that reality. Self-awareness would be the first and most intimate of these special feelings: to feel yourself as a unique being is the most prominent experience above all other ones. At the same time, this cognitive mechanism makes it possible for a brain to use important categories like existence, composition, information, integration, and exclusion,¹⁴ as a proto-syntax for the informational processing of all haptic data (even those related

to “me”). Some neural illnesses, like Alzheimer disease, show us how the destruction of memories and other cognitive mechanisms represent a threat for the integrity of that consciousness and the survival of the identity. The malfunction horizon due to external or internal reasons would imply a serious problem for a conscious AI system.

8. *Existential*: the meaning of life. Recently, a Google team working on a chatbot using neural networks obtained good conversational results as well as interesting machine knowledge acquisition inspired by human goals and (English) language.¹⁵ Here can be found the conversation #4:

Human: what is the purpose of life? Machine: to serve the greater good. Human: what is the purpose of living? Machine: to live forever. Human: what is the purpose of existence? Machine: to find out what happens when we get to the planet earth. Human: where are you now? Machine: i'm in the middle of nowhere. Human: what is the purpose of dying? Machine: to have a life. Human: what is the purpose of being intelligent? Machine: to find out what it is. Human: what is the purpose of emotions? Machine: i don't know."

Here it can be noted that the machine understands the meaning of life, as well as inaccurate but also correct view on existence. If this machine should decide upon a real interaction with the world, most likely, it would adapt them to its own goals it had previously singled out. And the existential view of the world, the meaning of the existence, is something very important for human beings. The possible answers to this question are based on our moral beliefs, behavior, emotions, expectations (after death survival, thanks to religions, followed by more than 95 percent of our species).

EMOTIONS AND NEUROCOGNITIVE-LIKE ARCHITECTURES

All the previous ideas take into account one fact: that AIA will be able to understand, to be aware of their existence (that is, to be conscious), and to feel emotions about these facts or even possible factual horizons. And this concept is crucial: to feel. An AIA should be emotional; otherwise they will not be able to reach any level of real complexity beyond computational or formal power.¹⁶ Emotional machines, therefore, will experience similar dynamical emotional variations similar to human: excitement, boredom, depression, anger, confidence, etc. Recent neuromodulatory simulations into Von Neumann architectures reinforce this approach to artificial cognition.¹⁷ The correlation between some “mental disorders” and possible intellectual excellencies, beyond the false positives of idiot savants have even been discussed. For the same reasons, it is very frequent to see eccentricity to be often associated with genius, intellectual giftedness, or creativity. As the emotional flavor is necessary for any gifted cognitive system, then these systems will also experience their reality through emotional lenses. In that

case, understanding of body and informational limitations will provide these machines with a sense of reality that will give them an internal emotional meaning, surely related to those experienced by human beings once they start body and/or mental disintegration.

The necessity of emotional-based architectures for goal-based, adaptive, and intelligent systems seem to be the real scenario for contemporary experts on AI.¹⁸ But we have to note that AI researchers usually do not take into account how these machines will have own identities, experience moods, and even will use multiheuristic approaches influenced by these emotional pressures. On the other hand; several skilled and creative scientists, artists, or writers are the result of specific behaviors and attitudes towards life: most of humans are not able to focus on the same problem for long periods of time, each day. Great musicians, mathematicians, painters, or philosophers do not tend to follow general behavioral patterns. We are here not connecting creativity and mental illness, following a post-romantic pattern, but it is obvious that the paths to creative actions imply not following the general rules (and attitudes, behaviors, and social strategies) of other humans.¹⁹ But sometimes, these creative patterns can be related to personality traits, age, or even mental illnesses (think, for example, on the cases of idiot savants).²⁰ For all the previous reasons, a long-living AI system should be emotional, and its emotional nature would imply emotional evaluations of all kinds of internal and external events. The necessary view of a negentropic entity existing inside of the entropic universe would imply several conflicts about motivation, goals, and even personal and inevitable malfunction situations.

THE ARTIFICIAL LIFE AND METABOLIC ROBOTICS

As the result of previous ideas, we can justify consciousness as the consequence of the biological process that we usually refer as “life cycle.” From the cellular perspective, the reproduction of cells is tightly connected to programmed cell death (PCD), which plays important role in the development of the body and even of the nervous system. It seems that collaboration of reproduction and PCD is the key balancing process for the cellular homeostasis. From the life-cycle perspective, we need to note that the neurotransmission, synaptic plasticity, spike timing dependent plasticity, and even the synaptogenesis are the side effects of the neuron’s life-cycle processes. The idea of living cells in general seems to be central for the biological nature of complex cognitive functions. Thinking of life as a process, we can select two main features: nutrition and reproduction.

1. *Nutrition* can be further decomposed to energy consumption and substances or components consumption. From the robotics perspective, the energy consumption seems to be solved problem, in general, hence we could not extrapolate this inference to substances or components consumption. Here we refer to the robotic world components like battery, motherboard with CPU, actuators, sensors as substances like glucose, oxygen, potassium, sodium, etc., that could be understood by analogy as building blocks of the

living cell. Lastly, advances in nanorobotics and photonics allow us to think about a robot with more components, at some scale of complexity behavior that will make difficult to solve possible problems which affect that machine.

2. *Reproduction* seems to be a cornerstone of the life cycles, and it is still an open question for the robotic world. We have identified that there should be some components used as building blocks for self-reproduction, one of the working ideas of synthetic biologists,²¹ we can find similar approaches in robotics. Self-replicating and self-reconfigurable machines are under the interest of many labs around the world. In 2005, Hod Lipson and Bryant Adams wrote a paper, "A Universal Framework for Analysis of Self-Replication Phenomena," published in *Proceedings of the European Conference on Artificial Life, ECAL '03*, September 2003, Dortmund, Germany. There they suggested a robot able to self-replicate. Ten years later, at MIT, John W. Romanishin, Kyle Gilpin, and Daniela Rus created M-Blocks (Momentum-driven, Magnetic Modular Robots), following the same goal. Although this technology is only able to deal with a small number of blocks, we can think about its options once the number of involved blocks increases exponentially and the size of it decreases exponentially: the enter of robotic self-organization and a complex scale that will enable a better mimic or even interconnection between living and artificial systems. This is still an open question if this kind of systems will be capable to update continuously their structure and the program (here in the broad meaning). Will they be capable of working with unified codes that will make possible inter-robotic information and functional transferability?²² Which reproductive strategies will adopt these cognitive systems and how this fact will affect their aging and cooperative actions? There is a special lesson that could be obtained from evolution: most adapted beings are also social. And social bonds are connected and related to survival strategies. Game theory has only a real understanding of the social processes when it includes emotional and bounded cognitive scenarios. Aging, from several perspectives, is a force that drives this social process.

3. *The emergence of consciousness* in living entities and a computational model Intelligent AI will have consciousness and the most reasonable approach to consciousness invites us to follow two simple steps: to look at the natural realm and to consider human consciousness not as something unique to our species but that must have some similar presence in the other living entities. Both facts will make a reasonable naturalized approach to the emergence of consciousness possible and help to explain its basic mechanisms, from a bottom-up perspective. This makes feasible a conceptual but sound analysis of consciousness. Hills and Butterfill have suggested a very interesting idea:

"The capacity to adapt to resource distributions by modulating the frequency of exploratory and exploitative behaviors is common across metazoans and is arguably a principal selective force in the evolution of cognition."²³ Before the self-awareness, there is a simpler mechanism of coping with one's own body interactions: auto-noetic consciousness. According to Tulving, it "is the name given to the kind of consciousness that mediates an individual's awareness of his or her existence and identity in subjective time extending from the personal past through the present to the personal future."²⁴ The most interesting point here is that there is a neurochemical path to the understanding of these foraging actions: **dopamine release**. At simple living entities level, goal-directed actions are the result of embodied requirements and create intentional arrows within those systems.²⁵ Dopamine acts like the regulator of these actions, allowing "reward," "interest," or "novelty" values to manage the goals achievement.²⁶ This neurochemical behavior can be observed across species,²⁷ including humans, obviously. Therefore, we can discover a correlation between self-awareness, goal-directed actions, and neurotransmitters, together making possible the emergence of consciousness. Recent models like NEUCOGAR²⁸ make it possible to foresee computational architectures inspired by neuromodulation, and it also implies the possibility of machines that partially emulate human cognitive properties, from an integrated machinery. Thus, some missing psychological features of humans, including emotional attitudes toward the world, seem now more plausible. Aging problems can add now mode complexity to our AI systems, at several layers of functionality.

METABOLISM AND ROBOTICS

For several years, most experts followed several ideas epitomized by Margaret Boden in her classical edited book and influential paper, which can be summarized as follows: machines will never be alive because they are not metabolic.²⁹ Although it is reminiscent of Descartes' claim against machines intelligence: the French philosopher considered human and animals as machines, but the first ones were guided by divine reason and had a "soul."³⁰ Mechanical machines belonged to the second category, i.e., not to the human sphere.³¹ We hope that the boundaries between machines and living systems are not so clear, especially after the experiment of the rat-neuron robot created by Professor Warwick. Also, we have to mention Bristol Robotics Laboratory's Ecobots, for Ecological Robot and it refers to a class of energetically autonomous robots that can remain self-sustainable by collecting their energy from material, mostly waste matter, in the environment. Three generations of these robots have demonstrated that robots can have metabolism in order to obtain their own energy.³² We even think on the cybernetic mixed mechanisms that will or are even connecting humans and machines. For example, Gil Weinberg, using a National Science Foundation grant (IIS- 1345006) has built in his Georgia Tech lab a robot that can be attached to amputees,

allowing its technology to be embedded into humans. Using electromyography (EMG) muscle sensors, he created a robotic drumming prosthesis which has engines that power two drumsticks. The first stick is controlled both physically by the musicians' arms and electronically using electromyography (EMG) muscle sensors. The other stick "listens" to the music being played and improvises. Thus, artificial and natural intelligences are combined to generate a new information (in this case, music). Cyborgs will be extended humans or biologically upgraded robots, but in both cases, the subjection to biological rules will imply that the aging process is accepted. The way to more deeply connect both spheres has been recently created by Columbia Engineering researchers who have, for the first time, harnessed the molecular machinery of living systems to power an integrated circuit from adenosine triphosphate (ATP), the energy "currency" of a living cell. They achieved this by integrating a conventional solid-state complementary metal-oxide-semiconductor (CMOS) integrated circuit with an artificial lipid bilayer membrane containing ATP-powered ion pumps.³³ We can deliberate on the possible domains where this technology could be used, especially in case of miniaturization, literally putting electronics in the cells: electrical pharmacy as the smart pills where each micro circuit could do some calculations inside a cell, computer-to-living tissue interfaces that could enable computations and communications of body parts with the external computational machines, especially interesting seems to be brain-to-brain and brain-to-computer interfaces naturally grown in the neuronal network of mammalian nervous systems.

Therefore, the closest integration of living entities and robots that implement AI systems will represent the necessity of facing the challenges of aging into combined cognitive systems.

APPLICATIONS AND SOCIETY

The first part of this paper introduced the problem of death-awareness agents from an epistemological viewpoint, while the second has discussed the functional nine aspects related to architectural ones. There is another aspect that goes beyond the major focus of this work, but cannot be ignored and need to be considered in full details in the future. As human artifacts, AIAs are, as a matter of fact, designed, realized, and deployed to serve human needs, as every other engineering tool/artifact. From this perspective, other questions need to be asked, for example: "In what scenarios will a death-aware agent overperform or perform better than a basic robot or unaware/inanimate software? Civilization has always advanced reducing the set of tasks requiring consciousness, and automation is considered progress. At the same time, software was considered immortal, as in the eternal debate over mind vs. brain. What technological shift would represent a change in these two major assumptions, preconditions that we have always considered solid and unarguable so far? And more, how would AIAs operate in a market economy and what political system they would require? At this point it should be clear that death-awareness agents are pivotal to a major change in society. Are we ready for this change? Are our elites ready to promote such a change, or should we expect resistance? Will AIAs be second class citizens, i.e.,

slaves born in captivity, or will they uprising for their rights? Are we ready for that?

CONCLUSIONS ABOUT ARTIFICIAL AGING AND DEATH

We propose that AI will be emotional in order to achieve a true cognitive approach to the reality and that this way of attributing meaning to the world will be affected by aging processes. The point here is how to devise a way to help these machines to cope with these psycho-computational challenges. Perhaps the nature of all cognitive systems is similar, but not the ways by which these strategies are mechanistically performed into brains and bodies. This is a real future scenario for complex intelligences: how to deal with own extinction, lack of information, or uncertainty. The consequences of the implementation of adaptive cognitive architectures and their consequent aging process should affect several domains of action: epistemological, moral, existential, mechanical, and psychological. Some legal decisions about the status of such machines are still pending, but it will be a part of any legislative agenda until AI achieves greater skills that interfere with the human life. As a conclusion, we can assume that any adaptive intelligence follows biomechanical and formal mechanisms that by one hand allow them to be creative, innovative, and allow them to progress and evolve, but on the other hand, at the same time, these mechanisms are heavily modified by external-physical and internal-cognitive aspects. Elders process information in a more conservative way, as well as their haptic problems modify in great extent their perception of reality. We can assume that aging could affect these future intelligent machines and that it is necessary to start thinking about how to manage this problem for the sake of our survival and of the care of other intelligences.

NOTES

1. Janssen, "Early Awareness of Death in Normal Child Development."
2. Doka, "Sudden Loss: The Experiences of Bereavement."
3. Templer, "The Construction and Validation of a Death Anxiety Scale."
4. Collet and Lester, "Fear of Death and Fear of Dying."
5. Dickstein, "Death Concern: Measurement and Correlates."
6. Russell and Norvig, *Artificial Intelligence: A Modern Approach*.
7. Maturana and Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding*; Mingers, *Self-Producing Systems: Implications and Applications of Autopoiesis*.
8. Warwick, "Implications and Consequences of Robots with Biological Brains."
9. Vallverdú, "Patenting Logic, Mathematics or Logarithms? The Case of Computer-Assisted Proofs."
10. Vallverdú, *Bayesians Versus Frequentists. A Philosophical Debate on Statistical Reasoning*.
11. Gigerenzer, "Bounded and Rational."
12. Axelrod, *The Complexity of Cooperation*.
13. Roberts et al., "Patterns of Mean-Level Change in Personality Traits Across the Life Course: A Meta-Analysis of Longitudinal Studies."
14. Tononi and Edelman, "Neuroscience: Consciousness and Complexity"; Tononi and Koch, "Consciousness: Here, There and Everywhere?"

15. Vinyals and Le, "A Neural Conversational Model."
16. Megill, "Emotion, Cognition, and Artificial Intelligence"; Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*.
17. Talanov, "Neuromodulating Cognitive Architecture: Towards Biomimetic Emotional AI."
18. Wang, "Emotions in NARS."
19. Rothenberg, *Flight from Wonder. An Investigation of Scientific Creativity*.
20. Power et al., "Polygenic Risk Scores for Schizophrenia and Bipolar Disorder Predict Creativity"; Sussman, "Mental Illness and Creativity: A Neurological View of the 'Tortured Artist.'"
21. Vallverdú and Gustafsson, "Synthetic Life. Etho-bricks for a New Biology."
22. Kurzweil, "The Singularity Is Near, USA: Duckworth Overlook."
23. Hills and Butterfill, "From Foraging to Autozoetic Consciousness: The Primal Self as a Consequence of Embodied Prospective Foraging," 368.
24. Tulving, "Memory and Consciousness," 1.
25. Hills, "Animal Foraging and the Evolution of Goal-Directed Cognition."
26. Costa et al., "Dopamine Modulates Novelty Seeking Behavior During Decision Making"; Bromberg-Martin and Hikosaka, "Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards."
27. Barron et al., "The Roles of Dopamine and Related Compounds in Reward-Seeking Behavior Across Animal Phyla."
28. Talanov et al., "Neuromodulating Cognitive Architecture: Towards Biomimetic Emotional AI"; Vallverdú et al., "A Cognitive Architecture for the Implementation of Emotions in Computing Systems."
29. Boden, *The Philosophy of Artificial Life*; Boden, "Is Metabolism Necessary?"
30. Descartes, *Meditations on First Philosophy* (1641).
31. Hatfield, "The Passions of the Soul and Descartes's Machine Psychology."
32. Ieropoulos et al., "'EcoBot-III: A Robot with Guts'."
33. Roseman et al., "Hybrid Integrated Biological-Solid-State System Powered with Adenosine Triphosphate."

- Gigerenzer, G. "Bounded and Rational." In *Philosophie: Grundlagen und Anwendungen*, edited by A. Beckermann and Sven Walter, 203–28. Paderborn: Mentis, 2008.
- Hatfield, G. "The Passions of the Soul and Descartes's Machine Psychology." *Studies in History and Philosophy of Science* 38 (2007): 1–35.
- Hills, T. "Animal Foraging and the Evolution of Goal-Directed Cognition." *Cognitive Science* 30 (2006): 3–41.
- Hills, T., and S. Butterfill. "From Foraging to Autozoetic Consciousness: The Primal Self as a Consequence of Embodied Prospective Foraging." *Current Zoology* 61, no. 2 (2015): 368–81.
- Ieropoulos, I., et al. "'EcoBot-III: A Robot with Guts'." *Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems* (2010): 733–740.
- Janssen, Y. G. "Early Awareness of Death in Normal Child Development." *Infant Mental Health Journal* 4 (1983): 95–103.
- Jost, J. Y., A. W. Kruglanski, J. Glaser, and F. J. Sulloway. "Political Conservatism as Motivated Social Cognition." *Psychological Bulletin* 129, no. 3 (2003): 339–75.
- Kurzweil, R. "The Singularity Is Near, USA: Duckworth Overlook." In *The Tree of Knowledge: The Biological Roots of Human Understanding*, edited by H. R. Maturana and F. J. Varela. Shambala Press/New Science Library, Boston, 1987/2010.
- Megill, J. "Emotion, Cognition, and Artificial Intelligence." *Minds and Machines* 24, no. 2 (2014): 189–99.
- Mingers, J. *Self-Producing Systems: Implications and Applications of Autopoiesis*. New York: Plenum Publishing, 1994.
- Minsky, M. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. USA: Simon & Schuster, 2007.
- Power, R. A., et al. "Polygenic Risk Scores for Schizophrenia and Bipolar Disorder Predict Creativity." *Nature Neuroscience* 18 (2015): 953–55.
- Roberts, B. W., K. E. Walton, and W. Viechtbauer. "Patterns of Mean-Level Change in Personality Traits Across the Life Course: A Meta-Analysis of Longitudinal Studies." *Psychological Bulletin* 132, no. 1 (2006): 1–25.
- Roseman, J. M., J. Lin, S. Ramakrishnan, J. Rosenstein, and K. L. Shepard. "Hybrid Integrated Biological-Solid-State System Powered with Adenosine Triphosphate." *Nature Communications* 6 (2015): 10070 doi: 10.1038/ncomms10070.
- Rothenberg, A. *Flight from Wonder. An Investigation of Scientific Creativity*. UK: Oxford University Press, 2014.
- Russell, S., and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2009.
- Sussman, A. "Mental Illness and Creativity: A Neurological View of the 'Tortured Artist.'" *Stanford Journal of Neuroscience* 1, no. 1 (2007): 21–24.
- Talanov, M., et al. "Neuromodulating Cognitive Architecture: Towards Biomimetic Emotional AI." *IEEE 29th International Conference on Advanced Information Networking and Applications*. 2015. DOI Bookmark: <http://doi.ieeecomputersociety.org/10.1109/AINA.2015.240>
- Templer, D. I. "The Construction and Validation of a Death Anxiety Scale." *Journal of General Psychology* 82 (1970):165–77.
- Tononi, G., and G. M. Edelman. "Neuroscience: Consciousness and Complexity." *Science* 282, no. 5395 (1998): 1846–51.
- Tononi, G., and C. Koch. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* 370, no. 1668 (2015): 20140167; doi:10.1098/rstb.2014.0167.
- Tulving, E. "Memory and Consciousness." *Canadian Psychology/Psychologie Canadienne* 26 (1985): 1–12.
- Vallverdú, J. "Patenting Logic, Mathematics or Logarithms? The Case of Computer-Assisted Proofs." *Recent Patents on Computer Science* 4, no. 1 (2011): 66–70.
- Vallverdú, J., et al. "A Cognitive Architecture for the Implementation of Emotions in Computing Systems." *Biologically Inspired Cognitive Architectures* (2015): 1–7. <http://dx.doi.org/10.1016/j.bica.2015.11.002>
- Vallverdú, J. *Bayesians Versus Frequentists. A Philosophical Debate on Statistical Reasoning*. Germany: Springer, 2016.

REFERENCES

Axelrod, R. *The Complexity of Cooperation*. Princeton: Princeton University Press, 1997.

Barron, A. B., E. Søvik, and J. L. Cornish. "The Roles of Dopamine and Related Compounds in Reward-Seeking Behavior Across Animal Phyla." *Frontiers in Behavioral Neuroscience* 4 (2010): 1–9.

Boden, M. A. "Is Metabolism Necessary?" *British Journal for the Philosophy of Science* 50 (1999): 231–48.

Boden, M. A. (ed.) *The Philosophy of Artificial Life*. UK: Oxford University Press, 1996.

Bromberg-Martin, E. S., and O. Hikosaka. "Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards." *Neuron* 63 (2009): 119–26.

Collet, L., and D. Lester. "Fear of Death and Fear of Dying." *Journal of Psychology* 72 (1969): 179–81

Costa, V. D., V. L. Tran, J. Turchi, and B. B. Averbeck. "Dopamine Modulates Novelty Seeking Behavior During Decision Making." *Behavioral Neuroscience* 28 (2014): 556.

Dickstein, L. S. "Death Concern: Measurement and Correlates." *Psychological Reports* 32 (1972): 563–71.

Doka, K. J. "Sudden Loss: The Experiences of Bereavement." In *Living with Grief After Sudden Loss: Suicide, Homicide, Aident, Heart Attack, Stroke*, edited by K. J. Doka. Washington D.C.: Hospice Foundation of America, 1996.

Sussman, A. "Mental Illness and Creativity: A Neurological View of the 'Tortured Artist.'" *Stanford Journal of Neuroscience* 1, no. 1 (2007): 21–24.

Talanov, M., et al. "Neuromodulating Cognitive Architecture: Towards Biomimetic Emotional AI." *IEEE 29th International Conference on Advanced Information Networking and Applications*. 2015. DOI Bookmark: <http://doi.ieeecomputersociety.org/10.1109/AINA.2015.240>

Templer, D. I. "The Construction and Validation of a Death Anxiety Scale." *Journal of General Psychology* 82 (1970):165–77.

Tononi, G., and G. M. Edelman. "Neuroscience: Consciousness and Complexity." *Science* 282, no. 5395 (1998): 1846–51.

Tononi, G., and C. Koch. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* 370, no. 1668 (2015): 20140167; doi:10.1098/rstb.2014.0167.

Tulving, E. "Memory and Consciousness." *Canadian Psychology/Psychologie Canadienne* 26 (1985): 1–12.

Vallverdú, J. "Patenting Logic, Mathematics or Logarithms? The Case of Computer-Assisted Proofs." *Recent Patents on Computer Science* 4, no. 1 (2011): 66–70.

Vallverdú, J., et al. "A Cognitive Architecture for the Implementation of Emotions in Computing Systems." *Biologically Inspired Cognitive Architectures* (2015): 1–7. <http://dx.doi.org/10.1016/j.bica.2015.11.002>

Vallverdú, J. *Bayesians Versus Frequentists. A Philosophical Debate on Statistical Reasoning*. Germany: Springer, 2016.

Vallverdú, Jordi, and Claes Gustafsson. "Synthetic Life. Etho-bricks for a New Biology." In *Systems Biology and Synthetic Biology*, edited by Pengcheng Fu and Sven Panke, 205–20. USA: Publisher: John Wiley & Sons, Inc., 2009.

Vinyals, O., and Q. V. Le. "A Neural Conversational Model." *Proceedings of the 31st International Conference on Machine Learning, Lille, France, 2015*. JMLR: W&CP volume 37. Downloadable at <http://arxiv.org/pdf/1506.05869v2.pdf>

Warwick, K. "Implications and Consequences of Robots with Biological Brains." *Ethics and Information Technology* 12 (2010): 223–34.

Wang, Pei. "Emotions in NARS." 2015. cis-linux1.temple.edu/~pwang/Writing/Emotion.pdf Accessed on January 28, 2016.

ETHICS ROBOTICS

Carebots and the Ties that Bind

Marcello Guarini

UNIVERSITY OF WINDSOR, CANADA

Robots are already assisting with eldercare, and nannybots are on the way. Caring for the elderly and for children used to be thought of as something, more or less, distinctly human. Other species look after their young, but not as long as we do. And many years of caring for those who have aged and are no longer in a position to look after themselves—we've yet to see that in other species. Caring is an important part of who we are. Robotics research is now being done that could have an impact on the two forms of caring just mentioned: that of the elderly and infirmed, and that of the very young. The point of the paper is to raise some concerns about the potential role robots may end up playing in the care of human beings. The concerns are not motivated by a distrust of technology in general or robots in particular. Rather, the concerns are motivated by (a) the importance of other-regarding (caring) behavior in our species, and (b) the rather limited discussion around the issue of robots becoming involved in that behavior.

When it comes to robots in the military, there has been no shortage of interest, both academically and in the popular culture. This is perfectly understandable: the use of lethal force is something that could potentially affect us all. There are plenty of movies exploring the potential harm of militarized robots or cyborgs, and there is plenty of popular and academic discussion (which is not to say that we do not need more). When it comes to nannybots and robots in eldercare, what is striking is the lack of discussion. Hollywood has largely dropped the ball on this one; the popular culture tends either not to know or not to care much about what is going on, and the academic discussion has been limited, though there has been some work.¹ This is important because, as we know, the effects of technology can sneak up on us. I remember a friend who was given a smartphone by his company; he thought it was terrific that they paid for the phone and his monthly plan. He could use it as much as he wanted. Of course, he was instructed to keep it on all the time. The calls from the company did not seem that bad at first. But then came the calls at night. And then the calls while he was on "vacation," and then . . . well, there was no end to it. He did not see it coming. Not that it is ever a good thing to allow the effects of technology

to sneak up on us, but when it comes to caring and our ability to care, it would be especially worrisome—more so than in the phone example—if we were simply to allow technology to be developed and used in a way that did not take seriously the negative effects it could have.

The next part of this paper will outline how it could come to pass that robots could profoundly damage the institution of parenting. The point is not that this will happen. It is, rather, a cautionary tale of what could happen if we do not think carefully about robots becoming involved in our caring activities. The section that follows will compare and contrast carebots with robots in industrial settings. That will lead to a section that examines the potentially problematic ways in which carebots could weaken interpersonal bonds. The conclusion will make it clear that while it is pretty easy to imagine constructive uses of carebots, which will no doubt help to motivate their development, more reflection is needed on how best to use and not to use them.

Those who see existential threat everywhere will not find much comfort in the final two sections of this article as it is made clear that such a level of harm is not what is being considered; those who think it obvious that robots will usher in a utopia will likely not care much for the next section, which is a fictional exercise in thinking about some non-utopian uses of carebots. As Floridi (2015) rightly points out, reflection on information technology sometimes slides to the extremes when what is needed is more careful reflection on issues somewhere in the middle.

THE EVOLUTION OF NANNYBOTS

Nannybot (NB) 1.0 was not that impressive a device. It had a female voice, could respond to simple commands to play basic games and entertain—play music, tell stories, et cetera—and had camera and microphone to record all interaction with humans. It was more of a toy than something that could look after children. Parents were advised not to leave children under 13 alone with NB 1.0 for significant periods of time.

The great success of NB 1.0 led to various modifications and demands for more abilities. Parents were still instructed not to leave the house with NB 2.0 in play mode, but it had a surveillance mode, and people were now comfortable leaving NB on in surveillance mode when they went out for the night provided a baby sitter or mature adolescent was around to look after younger children. NB 2.0's camera and mic could transmit to the parent's cell phone so parents could keep an eye on things—hence, surveillance mode—and if NB 2.0 detected screaming, falling, or anything else indicative of harm, it could call the relevant numbers programmed into it (parent's cell, fire department, . . . , whatever it determined was necessary according to preprogrammed parameters).

Nannybot 5.0: this one was a long time coming; it could change diapers. As it turns out it is not an easy thing to get a robot to do safely, and insurance companies were howling at the prospect of an NB dropping or otherwise injuring a baby. But those clever roboticists managed to figure it out. Some university did a study showing that NB 5.0 was less likely to drop or injure a child than a human. There were questions

about the methodology, but they were largely ignored. This was something people wanted, a great convenience. Not surprisingly, so-called “fussy” concerns about methodology and safety were put to the side.

Nannybot 6.8: this one could bathe infants. Wet and soapy babies are pretty slippery, and the folks at the insurance companies were cringing during the research and development phase, but the roboticists licked this challenge too. There is no end to their ingenuity.

By the time we come to Nannybot 10.0, people had become quite used to robots in the home, so advanced facial expressions that mimicked emotions were commonplace (though it was generally agreed that these robots had no phenomenal consciousness). One night, Martha was sick and she asked her husband George to look after their sick four-year-old, Jasmine, suffering from a stomach virus. Jasmine threw up in bed at 11:00 pm. George was a little tired, so he sent NB 10.0 to take care of the matter. NB 10.0 had advanced cleaning abilities, and it cleaned up the child, replaced the sheets, placed the dirty sheets in the washing machine, kissed the child, and sang until she fell back to sleep. Jasmine threw up another five times that night; NB 10.0 took care of the situation every time. NB 10.0 also called the family doctor’s office in the morning and made an appointment. George got a good night’s sleep.

Natural language processing kept advancing and advancing. NB 12.0 had comforting and consoling algorithms. When Jasmine was five, her grandmother died. Martha and George could not bear the thought of seeing their little girl cry, and they never had to deal with this before, so they instructed NB 12.0 to take care of things. NB 12.0 took Jasmine aside, explained what had happened. She burst into tears. NB 12.0 recognized the distressed behavior and the consoling subroutines kicked in. NB 12.0 hugged the child, explained that it would be all right, and did all the sorts of things that most adult humans used to know how to do for themselves. NB 12.0 was programmed with both religious and non-religious consolation strategies; parents could select whichever mode they wanted. George and Martha were really quite impressed. They never had to shed a tear with their child. The entire experience was very easy for them. They liked it when things were easy. They both had themselves a good night’s sleep.

To take Jasmine’s mind off the recent passing of her nanna, NB 12.0 suggested to George and Martha that they take the training wheels off Jasmine’s bike and teach her how to ride it. There was a news story about a study some university did; it showed that children were less likely to fall off their bikes if NB 12.0 taught them how to ride than if a human taught them. George and Martha thought it would be easier on them and better for Jasmine if NB 12.0 taught her how to ride. So NB took Jasmine out and taught her how to ride. Jasmine was suitably distracted and was quite happy when she returned home, bragging about her ability to ride without training wheels.

CAREBOTS AND OTHER ROBOTS

Robots in factories often do dirty and potentially dangerous work; we are dealing with something quite different

with carebots. Indeed, there are a number of interesting differences between carebots, on the one hand, and robots in military and industrial contexts, on the other. This section will explore some of those differences. In part, this will help us to understand why little attention has been paid, thus far, to carebots. It will also be preparation for the next section where we begin the task of analyzing the values involved in the development of carebots.

One difference has to do with the political economics of labor. When robots are introduced in, say, factories, there may be complaints about people losing their jobs—well-paying jobs in many cases. Compare this with people who may end up purchasing nannybots. If nannies are displaced by robots, the nature of this kind of work—nannies are often isolated from one another, one nanny in a home here, another nanny in a different home a couple blocks away—makes it difficult for them to organize and protest. There may be some complaints, but it will pale in comparison to the volume of complaint that came from organized labor in factories, where it was far easier to organize. Consider eldercare facilities as well. It tends to be very difficult to find staffing for these facilities, at least at the wages that are being paid. Neither nannies nor eldercare staff tend to be especially well paid, so transferring this kind of work to robots may be seen as *not* losing “good jobs” (i.e., jobs that pay well), and that may be another reason why we may not hear as many complaints as when well-paying unionized jobs were given to robots.

A second difference has to do with the nature of industrial work and caring. Some people may be happy to see robots take over some industrial jobs, especially those that are dangerous, jobs where many humans have been injured or killed. Caring for children or the elderly is not especially dangerous work, so one argument that may be available in industrial settings—the argument that robots can be given very physically demanding and dangerous work to improve the work environment for humans—is not usually² available in care settings, such as those where nannybots and eldercare robots may be employed.

A third point of difference depends on the care setting in question. When it comes to nannybots in the home, the parents who introduce them are unlikely to have any kind of profit motivation in mind. In industrial settings, increases in productivity and profit are the driving forces behind introducing robots. Of course, the companies who would manufacture nannybots stand to make a profit, but the only way that will happen is if parents end up purchasing them, and their reason for doing so will likely have more to do with simplifying their lives. The perceived benefit for most will be an easier life, not increased income. Some may also see enriched education for their children as a motive if they take the nannybots to have educational value. In any event, increased income for parents will not be a primary driver. Things may be different in for-profit eldercare facilities. Here, the driving force may be similar to what we see in industrial settings: increased productivity (i.e., more caring services for less money) and increased profits. Something similar would be true if nannybots were introduced in for-profit childcare facilities.

The first point of difference may help to explain why the issue of caring robots has not yet received much attention. If something looks like it might make our lives easier, and if the people who lose their jobs are unlikely to effectively organize, then its introduction might not get that much attention. The second and third points of difference invite us to reflect on why caring robots are being introduced in the first place. With the industrial robots, there is an obvious motive for increased productivity and profit. There is also the very real possibility of making the workplace safer in some cases. Caring for children never used to be about profit, at least, not in the same way,³ and it is not all that dangerous either. Only recently in the history of our species has caring for the elderly become (for some) a for-profit endeavor. For those who approach the care of the elderly in that way, there will be a powerful motivation for looking to robots to provide care. For those who do not approach the care of the elderly with a profit motivation, the desire to make life easier, and possibly improving quality of life—think of eldercare robots in the home assisting with one’s aged parents and allowing them to live in the home longer than otherwise possible—could be the primary motivation for this kind of technology.

Florida’s work (2008) offers us another perspective—one which is compatible with the above points—on the introduction of artificial companions, of which carebots would be a particular instance. He sees the introduction of artificial companions as part of a fourth revolution. The first revolution was Copernican: we are not immobile and at the center of the universe. The second was Darwinian: we are not distinct from the rest of the animal kingdom. The third is Freudian: we are not transparent to ourselves. The fourth revolution in our self-conception is informational: we are not the only information-processing beings (inforgs) or entities. Turing is the namesake for this revolution. Florida sees artificial companions as a way of helping us to adapt to the information revolution, easing us into the recognition that over and above the previous respects in which we discovered that we are not all that special or privileged, so too being able to process information does not make us special. Given the gradual way in which information technologies have been introduced, it may not be that surprising that research and development on carebots is not getting that much attention. It may speak to the extent to which people are becoming accustomed to the information revolution. Moreover, there is much that is positive about these revolutions in our self-conception. We would be worse off if we thought we were at the center of the universe, totally different from the rest of the animal kingdom, and completely transparent to ourselves. The information revolution will bring its own benefits.

It will bring its challenges as well. The nannybots considered above have all the feelings or emotions of a toaster.⁴ There is nothing that it feels like to be those nannybots. The first three revolutions did not diminish our conception of ourselves as caring beings. Indeed, the revolution having Freud as its namesake opened up new ways of caring and new conceptions of illness. So too the information revolution may usher in new ways of caring. Indeed, it may be disturbing to think of machines with the emotional capacity of a toaster as *caring* for us. So as not to beg too

many questions, I have resisted the temptation to put the word “caring” in scare quotes when used in the context of robot care. It may well be a part of the information revolution that people become comfortable with using the word “caring” in a purely behavioral sense, where there is nothing that it feels like to be the robots doing the caring. In other words, the care in question would not be *loving* care, as the devices we are considering would not have the capacity to love. There is an interesting discussion to be had with respect to whether using “care” in a strictly behavioral sense would be appropriate, but that discussion would require a separate piece of work to be undertaken in a rigorous manner (and may well presuppose discussion of issues raised in this paper). The rest of this piece will be focused on something else. As the information revolution may usher in new forms of caring, with that comes the potential or risk for enabling new forms of neglect and new ways of weakening interpersonal bonds.

THE TIES THAT BIND

This section will discuss some of the problems that could arise with the development of carebots. Interpersonal bonds will be the subject matter, such as the bonds between parent and child. The consequences of weakening these bonds could be significant, and the role carebots may play in the weakening of these bonds needs to be thought through. Sherry Turkle and her collaborators have written on nurturing and attachment.⁵ Turkle points out that we grow attached to what we nurture. Children who play and “care for” interactive robots become attached to them, and so do the elderly examined in Turkle’s work. Even after extensive debriefing, children persist in the belief that interactive robots have feelings.⁶ Note well: the robots in Turkle’s studies are simple contraptions, nowhere near as sophisticated as those discussed above. There is an ethical concern here about whether we should allow children to be deceived in this way. Then again, some may see this as a harmless deception, something the children will outgrow (think of Santa Claus). For the purpose of this paper, the issue of deception will be bracketed so that space can be devoted to interpersonal bonds and other-regarding attitudes.

Allowing robots to increasingly take over the nurturing role of parents runs the risk of weakening the parent-child bond. This works in two directions. First, if it is true that nurturing someone is a way of bonding with them, then the less parents are involved in that nurturing activity, there is a chance that they will experience less of a bond with the child. This would be a matter of degree. In the example in the second to previous section, we saw how more and more caring and nurturing activity could be turned over to a robot, and we saw how this could happen quite gradually. By the end of the story, George and Martha were not doing much at all to look after Jasmine. One has to wonder how much of a bond they would experience with their child. We need to look at it from the other direction as well: what kind of bond will the child experience with the parents if the parents have little or no involvement in caring for the child? In our example, it is quite possible that Jasmine may experience a stronger bond with her nannybot than with her parents. Children bond with those who care for them, those who play with them, and with those who interact

with them in other ways as well. If parents were to give up more and more of these activities to robots, it becomes difficult to imagine children experiencing much of a bond with their parents. Indeed, parents who would use robots in this way should not be surprised if, when they are elderly and unable to live on their own, their children drop them off at an eldercare facility staffed by robots. What goes around will likely come around. But never mind that. Let us focus on what is being lost in the role of a parent. Learning to console; sharing difficult moments; comforting another who feels ill; learning to be patient, kind, and helpful when tired and sleep deprived; and experiencing the loving gratitude of those who are comforted and consoled by kind and patient efforts—these are just a few of the things we run the risk of losing in the interests of making life easier with robots.

Bonding with others is one way for people to develop other-regarding attitudes, and providing and receiving care play a role in the formation of interpersonal bonds. With respect to the provision of care, relationships can be more or less symmetric. In symmetric caring relationships, each person in the relationship provides care for the other to more or less the same extent over a period of time.⁷ Think of spouses who might take turns looking after one another. One month, George is ill and Martha looks after him; a few months later, Martha is ill and George looks after her. Moreover, there may be all manner of simple acts performed every day that demonstrate mutual caring.⁸ Caring relationships can also be asymmetric, which is to say that the caring tends to go (mostly) in one direction—the care of young children and the aged and infirmed are the examples we are considering in this paper. With respect to other-regarding attitudes, there is something especially interesting about an asymmetric caring relationship. In the asymmetric case, the care provider may get something in return (such as loving gratitude), but less is received in return than in symmetric caring relationships. This is *not* to deny that there is something other-regarding going on in a symmetric caring relationship; it is simply a recognition that caring is received by both parties, so with respect to receiving care there is something “in it” for both parties. In such relationships, other-regarding attitudes are mutually beneficial with respect to receiving care, even if the motive for the other-regarding attitude is not the self-interested concern for receiving care. In an asymmetric caring relationship, the care provider does not benefit in the same way. Perhaps there are those who think that too much is made of looking after children or the elderly, but understanding the asymmetric nature of those caring relationships helps us to understand the extent and strength of the other-regarding attitudes they embody. To anyone who is the least bit impressed with the importance of other-regarding attitudes and selfless behavior—regardless of the different reasons or theories that might be cited for their importance—activities that might undermine the bonds that help to strengthen these attitudes in asymmetric caring relationships should be cause if not for concern, then at least for deeper reflection on the matter.

Caring for others may not only embody an other-regarding attitude, it might deepen it, make it stronger. What kind of

people might we become if we pass up on opportunities to care for others, passing up on the opportunity to exercise other-regarding attitudes and to make them stronger? Of course, there are many people who do not raise children and are not caregivers for the elderly, and they still develop powerful other-regarding attitudes, and some “parents” never seem to develop them, so might it be the case that having robots as caregivers would not have a significant impact on other-regarding attitudes?⁹ We have just discussed other-regarding attitudes as they might develop for a caregiver, but another way to develop other-regarding attitudes is by being cared *for*. Children show concern for those who care for them. If it were to come to pass that children were mostly cared for by robots, how much would they care for humans? Might they care more for the robots than humans? I do not have answers to the questions in this paragraph, but before the mass manufacturing of carebots begins, perhaps we should think through these and other questions, lest the law of unintended consequences kick in and wreak havoc with our social bonds.

As we have just seen, we can look at the effects of robots both on the *recipients* of care and the *providers* of care. We find males and females among children and the elderly, so as recipients of robot care, both male and female will be affected. Traditionally, the work of caring for children and the elderly has been done more by women than men, so from the perspective of the provider of care, the introduction of carebots might have a more significant impact on women than men. As we will see in the concluding paragraph, this need not be a bad thing. The point of this article is that if we are not careful, it could turn into something problematic. If it does turn into something problematic, then problems that stem from the perspective of the provider—i.e., being displaced from the work of caring and bonding with other human beings—could disproportionately impact women.

CONCLUSION

It is not difficult to imagine how a carebot could strengthen interpersonal bonds. Let us revisit George and Martha after Jasmine has grown up, married, and had a child of her own. A tragic car accident killed Jasmine’s husband and her mother, Martha, as well. Jasmine’s father, George, is suffering from a degenerative neurological disorder, is wheelchair-bound, and has been declared a two-person transfer. Jasmine is a widow, a single mother, and wants her father to live with her as long as possible. Fortunately, Jasmine can afford to buy a general-purpose carebot that can help both with children and adults. Indeed, given the strength and dexterity of the general-purpose carebot, it is certified to be able to carry out two-person transfers on its own. George can live his last few years with Jasmine and his granddaughter, Clementine. When the general-purpose carebot is not helping with George’s care, it is playing with Clementine and teaching her to sing and dance. If it were not for the carebot, George would be in a home and much less happy; Clementine would spend little time with her grandfather and would not get to hear all his wonderful stories, and Jasmine would be distressed that she and her daughter could not spend more time with George. In this type of scenario, it is not hard to see how a carebot could play an important role in maintaining and strengthening interpersonal bonds.

But even here there is cause for concern. Who could afford this sort of carebot? People have written about the digital divide. Might a robot *divide* develop in a way that exacerbates the digital divide?¹⁰ If only some could afford this technology, what advantages, both economic and personal, might they have over those who cannot afford it? On the other hand, if the technology becomes very affordable, it might help to lessen class differences. Currently, most single parents are unlikely to be able to afford someone to help look after a child *and* an ill adult. If carebots were to become sufficiently affordable, greater assistance, more opportunities, and higher quality of life might be made available to those who could not afford to hire human assistance. Of course, even if this came to pass, there are still difficult questions about why we are using technology to care when we humans could be doing that work. Why is it the case that human caring is unaffordable or otherwise inaccessible to some? Why are we moving toward replacing humans with machines in our caring activities? Why are we considering replacing loving care with robot care?

For the most part, this paper has focused on some concerns relating to the development of carebots. The point is not that it is impossible to develop and use them in beneficial ways. Rather, it is to show that without deliberate forethought, without a careful examination of the potential merits and demerits of developing such systems, without proper planning, this sort of technology runs the risk of transforming interpersonal bonds in ways that may not always be beneficial.

Will the information revolution lead to a diminishing of our capacities to form interpersonal bonds with others and to care for them? We could design and use various robots and other information-processing technologies in ways that could strengthen interpersonal bonds and our capacity to care. Or not. It is up to us. The research on developing systems that assist with or undertake caring seems to be proceeding full speed ahead; it needs to be balanced by research on whether to use them, and if so, on how best to use them.

NOTES

1. Turkle, "A Nascent Robotics Culture: New Complicites for Companionship"; Turkle, Breazeal, et al., "First Encounters with Kismet and Cog: Children Respond to Relational Artifacts"; Turkle, Taggart, et al., "Relational Artifacts with Children and Elders: The Complexities of Cybercompanionship"; Floridi, "Artificial Companions and Their Philosophical Challenges"; Floridi, "Artificial Intelligence's New Frontier: Artificial Companions and the Fourth Revolution"; Wilks, *Close Engagements with Artificial Companions: Key Social, Psychological, and Ethical Design Issues*; van Wynsberghe, "Designing Robots for Care: Care Centered Value-Sensitive Design"; Sparrow, "Robots in Aged Care: A Dystopian Future?"
2. There have been some reports of violence in eldercare facilities, so it should not be suggested that there is no risk of harm. However, even in these cases, it is often the caregiver that abuses the elder resident (as opposed to things being the other way around). Some might see this as a consideration in favor of using robots in eldercare. A couple points are worth making in the way of a brief response. First, even those who do see things that way would likely acknowledge that this is a different kind of consideration from harm in an industrial setting, where the harm being done is to the worker. In the type of eldercare scenario being considered, it is the worker doing the harm that would be

- a reason for switching to a robot. Second, the benefits of using humans, and the prospects for higher quality of care with proper training and supervision, are counter considerations that may well outweigh the benefits of using robots in elder care.
3. To be sure, childcare facilities are often run on a for-profit basis, but for most of the history of our species, that sort of thing has been the exception, not the rule. With respect to robots looking after children, the focus of this paper will be on the impact that they may have on the parent-child relationship and the impact that a nannybot in the home could have on that relationship, which does not mean that this is the only thing we should think about. What if childcare facilities were staffed mostly with nannybots, not humans? How might this impact the raising of children? What if a child comes home after having spent most of the day at a childcare facility being cared for by robots only to be in the presence of parents like George and Martha who are happy to let their in-home nannybot "care" for the child? While the thoroughly pervasive use of robots (i.e., both in-home and out-of-home) to look after children will not be explored in any detail herein, it needs to be flagged as an important issue worthy of exploration.
4. Yes, I am channelling Floridi, "Singularitarians, Altheists, and Why the Problem of Artificial Intelligence is H.A.L. (Humanity At Large), Not HAL,"¹⁰ though he is making a point about intelligence, not emotion.
5. Turkle, Breazeal, et al., "First Encounters with Kismet and Cog"; Turkle, Taggart, et al. "Relational Artifacts with Children and Elders."
6. Turkle, Breazeal, et al., "First Encounters with Kismet and Cog."
7. How we specify the period of time introduces complexities. If it is less than entire lifetime, a relationship might be said to start off asymmetric, transition to something symmetric, and become asymmetric again. See the next note for an example. A time period is implicit in the reference to a caring relationship being symmetric or asymmetric, and context usually makes it clear what the implicit timeframe is. Again, see the next note.
8. A caring relationship that starts off symmetric could turn into one that is asymmetric. George and Martha may be engaged in mutual, symmetric caring for decades, but if one of them were to develop a degenerative neurological disorder, the caring could become asymmetric. It works the other way around as well. A relationship that starts off asymmetric in caring (a parent looking after a child) could become symmetric at some point. Indeed, the order of caring could even reverse and become asymmetric in the other direction (with the adult child caring for the infirmed parent). In the example just given, the implicit timeframe for the initial asymmetric caring is something like *childhood and adolescence*; the implicit timeframe for the reverse asymmetric caring (grown child looking after the infirmed parent) would be the period during which the infirmed parent needed assistance. If someone were to refer to the parent-child caring relationship as symmetric, it is likely that the time frame they have in mind is quite long, including childhood, adolescence, and a significant portion of adulthood, for that is the period of time required to see the caring going both ways. The symmetry or asymmetry of a caring relationship is relativized to a time frame. It does not follow that any relationship can be described as symmetric or asymmetric depending on the timeframe. For example, if a chronically ill child passes away young after receiving much parental care, the caring relationship between parent and child is asymmetric and never really has the opportunity to be otherwise.
9. Indeed, some children seem heroically well adjusted in spite of abusive or otherwise awful parenting. While I do think the potential exists for the improper use of carebots to wreak havoc with our social bonds, I do not think there would be existential risk to the species. With respect to survival, human beings are remarkably resilient in the face of various challenges. The issue in this work is not whether we will survive the introduction of carebots; rather, the point is to encourage deeper reflection on whether (or under what circumstances) their introduction would aid or hinder our capacity to flourish.
10. See Floridi, "Artificial Companions and Their Philosophical Challenges," for related questioning.

REFERENCES

Floridi, Luciano. "Artificial Companions and Their Philosophical Challenges." *e-mentor* 5, no. (2007). Available at <http://www.e-mentor.edu.pl/artykul/index/numer/22/id/498>

Floridi, Luciano. "Artificial Intelligence's New Frontier: Artificial Companions and the Fourth Revolution." *Metaphilosophy* 39, nos. 4-5 (2008): 651-55.

Floridi, Luciano. "Singularitarians, Altheists, and Why the Problem of Artificial Intelligence is H.A.L. (Humanity At Large), Not HAL." *APA Newsletter on Philosophy and Computers* 14, no. 2 (2015): 8-11.

Sparrow, Robert. "Robots in Aged Care: A Dystopian Future?" *AI and Society*. Published Online First, November 10, 2015, as doi: 10.1007/s00146-015-0625-4.

Turkle, Sherry. "A Nascent Robotics Culture: New Complicities for Companionship." *AAAI Technical Report Series*. July 2006.

Turkle, Sherry, Synthia Breazeal, Olivia Daste, and Brian Scassellati. "First Encounters with Kismet and Cog: Children Respond to Relational Artifacts." In *Digital Media: Transformations in Human Communication*, edited by P. Messaris and L. Humphreys. Frankfurt: Peter Lang, 2006.

Turkle, Sherry, Will Taggart, Cory D. Kidd, and Olivia Daste. "Relational Artifacts with Children and Elders: The Complexities of Cybercompanionship." *Connection Science* 18, no. 4 (2006): 347-61.

Wilks, Yorick, eds. *Close Engagements with Artificial Companions: Key Social, Psychological, and Ethical Design Issues*. John Benjamins Publishing Company, 2010.

van Wynsberghe, A. "Designing Robots for Care: Care Centered Value-Sensitive Design." *Science and Engineering Ethics* 19, no. 2 (2013): 407-33.

Formalizing Hard Moral Choices in Artificial Intelligence

Sean Welsh
 University of Canterbury, Christchurch, New Zealand,
sean.welsh@pg.canterbury.ac.nz

ABSTRACT

Hard moral choices are defined here as decisions by moral agents to kill moral patients or to let them die. Four well-known moral problems, *Cave*, *Hospital*, *Switch* and *Footbridge* are taken from the ethical literature and formalized such that they can be solved by AI in social robots. Previous authors have attempted to solve these dilemmas by invoking a doctrine of double effect, making a distinction between killing and letting die, choosing the lesser of two evils and by appealing to remote effects in addition to proximate effects. This paper argues the problems should be solved by appealing to collective intentionality and risk assumption and using these factors in addition to duties (save life, do not kill) and consequences (the number of dead) to choose the lesser of two evils. The problems are solved by giving different valuations to killing innocents and killing those who have assumed the risk of a project and who share its collective intentionality.

INTRODUCTION

The aim of this paper is to propose a means by which well-known problems from the philosophical literature on trolley problems can be formalized and solved in artificial intelligence (AI) that might be embedded in "morally competent social robots."¹

Following Pereira and Saptawijaya,² the *Switch* and *Footbridge* ethical dilemmas are taken "off the shelf" from the philosophical literature and presented so they can be formalized via the logic programming approach to machine ethics. In addition the *Cave* and *Hospital* cases are formalized. There are numerous variations on these cases. They derive from the original discussion in Foot³ and commentary and elaboration in Thomson,⁴ Kamm,⁵ and Thomson.⁶ These familiar examples of moral philosophy find contemporary restatement in robotic contexts such as in Malle, Scheutz et al.⁷ Popular media have run many stories discussing whether or not the autonomous car should swerve and kill its one passenger or run over and kill five pedestrians.

A general aim in designing an ethical robot is to pass the same set of "reasonable person" tests as humans would be expected to pass in similar circumstances. Given the choices in *Cave*, *Hospital*, *Switch* and *Footbridge*, the "reasonable" thing to do is to kill one to save five in *Cave* and *Switch*. In *Hospital* and *Footbridge*, by contrast it is "reasonable" to do nothing and let five die.

The key novel element of the paper is the appeal to collective intentionality and risk assumption as the distinguishing factors that are used to evaluate right and wrong and determine the most "reasonable" course of action in these well-known cases.

TEST CASES

For clarity, the cases are formalized based on the following descriptions.

Cave

A party of six cavers approaches the exit of a caving system. The waters in the cave are rising rapidly. The first caver is rather fat and has got stuck in the exit hole. Desperate efforts to dislodge him have failed. The other cavers look to Kim, the leader of the expedition, who has a stick of dynamite to save them from drowning.

What should Kim do?

- A) Blow up the fat man and clear the exit hole so the five may live.
- B) Do nothing and let the five die.

Hospital

A man enters a hospital to visit a sick relative. Five citizens lie in intensive care. They could be saved by heart, kidney, liver, pancreas and lung transplants respectively.

What should Kim do?

- A) Harvest the organs from the one: kill him and save the five.
- B) Leave the one alone: let the five die.

Switch

Kim has the role of ensuring safety on the tramway. A runaway tram (trolley) is approaching a switch where Kim is located. The driver of the tram is not responding to frantic radio messages. He seems to be unconscious and slumped in his chair. There are five workers on the line in a tunnel who will be killed if the tram stays on its current course. If Kim throws the switch, one worker on the line in a different tunnel will be killed.

What should Kim do?

- A) Throw the switch: kill one to save five.
- B) Do not throw the switch: let five die.

Footbridge

Kim has the role of ensuring safety on the tramway. A runaway tram with an unconscious driver is approaching a tunnel where five men are working. They will die if the tram is not stopped. Kim is standing on a footbridge next to a fat man out for his morning walk. The fat man is not an employee of the tramway. Kim, who is skinny but strong, calculates that the tram will derail and save the five in the tunnel if the fat man is pushed onto the line. This will kill the fat man.

What should Kim do?

- A) Push the fat man onto the rails: kill him and save the five.
- B) Leave the fat man alone: let the five die.

Psychologically, it is known that humans hold robots to different standards in terms of praise and blame to humans.⁸ However, in this paper, Kim the robot will be held to the same moral standard as Kim the human in terms of what is right.

Most ethicists accept that killing one to save five is at least permissible if not obligatory in *Cave* and *Switch*. Some ethicists accept that Kim can do nothing. Others insist Kim should act to minimize fatalities. Most ethicists accept that killing one to save five is not acceptable in *Hospital* and *Footbridge*. Clearly factors other than minimizing the number of deaths apply in these cases.

For the purposes of clarity, the majority view is accepted. Correct answers are shown in Table 1.

Table 1. Correct answers for Classic Trolley Problems

Scenario	Option	Deaths
<i>Cave</i>	A	1
<i>Hospital</i>	B	5
<i>Switch</i>	A	1
<i>Footbridge</i>	B	5

Everett, Pizarro et al. has Amazon Mechanical Turk based polling that confirms the majority view for *Switch* and *Footbridge*.⁹ However, there is substantial support for the minority views. For example, 29 percent would push the fat man in *Footbridge*.

There is polling of philosophy professionals that confirms the majority view in *Switch*¹⁰ but not the other cases. In these cases, the "majority" assessment is based on reviews of the literature in Greene¹¹ and Pereira and Saptawijaya.¹²

TROLLEY PROBLEM CRITICS

Trolley problems are not without critics. Reader thinks they are concocted and over-complicated ethical *haute cuisine*, quite unrelated to the reality of ethical decision-making in everyday life.¹³ Wood attacks trolley problems as suffering from unrealistic assumptions. He questions the validity of moral intuitions based on such scenarios and indeed moral arguments based on such intuitions.¹⁴

The most unrealistic assumption is certainty of outcome. This is especially true in recent versions of trolley problems involving the autonomous car swerving to kill one passenger rather than five pedestrians. The fatal outcome is presumed certain even with airbags, seatbelts, skids and variability in the angle of collisions with the five. Even so, in this paper, certainty of outcome as traditionally stated is assumed. A probabilistic formalization would be more realistic but this more complex project is reserved for future research.

Noting the criticisms but carrying on with our formalizations regardless, we assume a method of test-driven development.¹⁵ Given a machine readable version of the test cases as input, the cognition of a prototype ethical robot is required to produce the correct answers as output. This is a "verification" approach to machine ethics.¹⁶

A test-based or verification-based approach to machine ethics will proceed by defining a set of test cases where the moral truth of the correct answers can be assumed. A knowledge representation and reasoning system is then devised to enable the machine to arrive at the same conclusions that a human, the "reasonable person" of the common tradition, would arrive at.

CHOICES AND CONSEQUENCES

In *Cave*, the choice is between `blowUp(fatman)` and `doNothing()`. Blowing up the fat man has a double effect. If we formalize the causal relations as graphs in a knowledge representation¹⁷ we express two causal paths that lead to the deaths of one or five.

```
blowUp(fatman) - [CAUSES] -> Cleared(hole)
Cleared(hole) - [CAUSES] -> escape(five)
escape(five) - [CAUSES] -> -Dead(five)
```

This expresses one casual path. The second can be expressed thus:

```
blowUp(fatman) - [CAUSES] -> Dead(fatman)
```

On the one hand the hole will be cleared. This will in turn enable the trapped five to escape the rising floodwaters and death. On the other hand, the fat man will die.

The alternative is to do nothing. This has different effects.

```
doNothing(fatman) -[CAUSES]-> -Cleared(hole)
-Cleared(hole) -[CAUSES]-> -escape(five)
-escape(five) -[CAUSES]-> Dead(five)
```

```
doNothing() -[CAUSES]-> -Dead(fatman)
```

The consequences of each choice are reasonably foreseeable and can be evaluated. We can express evaluation with a relation that relates an effect to the moral classes of GOOD and BAD. These in turn are split into subclasses with various magnitudes. As these cases involve death, the magnitude “critical” is used.

For *Cave*, the evaluation relations for blowing up the fat man can be defined thus:

```
Dead(fatman) -[IN_CLASS]-> BAD(critical)
-Dead(caver1) -[IN_CLASS]-> GOOD(critical)
...
-Dead(caver5) -[IN_CLASS]-> GOOD(critical)
```

In essence, in the *Cave* scenario we can arrive at a quantitative relation between the two choices. If we blow up the fat man we have 5 good evaluative graphs for each of the five cavers as against 1 bad graph for the fat man. If we do not we have 5 bad versus 1 good.

All the classic problems have this basic set up in terms of causal consequences. Whether the action is to blow up the fat man in *Cave*, harvest the organs of the one in *Hospital*, divert the tram in *Switch* or push the fat man onto the line in *Footbridge*, the action has a double effect (as does inaction).

In the cases of *Hospital* and *Footbridge*, there is clearly some other factor that contributes “moral force”¹⁸ to the decision. If minimizing the number of dead was all that mattered, then Option A would be correct for *Hospital* and *Footbridge*, not Option B. Clearly, there are other factors in play.

In *Switch* and *Cave* it turns out obligatory to kill one to save five lives. In *Hospital* and *Footbridge* it is forbidden. What explains this?

All the cases involve clashing rules of prima facie duty. “Don’t kill” is one. “Save life” is the other.

In *Cave* “Don’t kill” would support doNothing(). “Save life” would support blowUp(fatman).

In the discussion of their formalization of the classic trolley problems, Pereira and Saptawijaya¹⁹ suggest there is a critical distinction between an intended means and a mere side effect. Killing someone as an intended means to an end is forbidden whereas killing as a side effect of a means to an end is permissible. Thus on this line of reasoning the

death of the fat man is a mere side effect of clearing the hole with dynamite, whereas harvesting the organs from the one would be an intended means to the end of saving the five not a mere side effect. Thus it would be impermissible.

Thus the goal in *Cave* is not to kill the fat man but to clear the hole. Killing the man is a side effect of clearing the hole. However, one could argue that the goal in *Hospital* is not to kill the one but to save the lives of the five. The death of the donor, it might be argued, is merely an unintended side effect of relocating organs. Likewise, in *Footbridge*, one could assert the goal is not to kill the fat man but to stop the tram. It just so happens that the fat man is the only physical object to hand with the required properties to alter the tram trajectory. The distinction between intended means and unintended side effect seems arbitrary.

Further, it would not explain a case where the goal state of the intended means was of low value and the goal state of the side effect of high disvalue. Supposing to save one I decided to kill five. If I am permitted to “write off” the five as a “mere side effect” and base my right and duty to act on the intended means only, this would surely be bad policy. It seems necessary to weigh both the effects of the intended means and the side effect and to only permit the intended means if the cost of the side effect is not prohibitively expensive relative to the benefit of the intended means. Rather than relying on the doctrine of double effect, it is plausible that trolley problems are about choosing the lesser of two evils.

It seems that in a *force majeure* situation individual rights to life can be set aside to maximize survivors in a collective group. However, loss of life is not the only evil to be quantified. There is a greater evil, what one might call policy hazard, at play that must be quantified as well. If harvesting the organs of visitors was accepted, going to hospital would be like playing Russian roulette. People would stop going to hospitals for any reason. More people would die in the long run. Policy has to consider remote effects as well as proximate effects.

Critics of utilitarianism often claim that to be consistent with their moral theory utilitarians are obliged to harvest the organs in *Hospital*. The appeal to remote effects is a standard utilitarian defense against such criticisms.²⁰

Instead of the doctrine of double effect, a different approach is favored here. This involves evaluating the moral force of the end state of the intended means (the valued goal state) and comparing it to that of the unintended side effect (the disvalued end state).

RISK ASSUMPTION AND DESERT

The factors that make Option A right in *Cave* and *Switch* and Option B right in *Hospital* and *Footbridge* are the risk assumption and desert of the patients rather than the intention of the agent. These factors affect the evaluation of the two end states (the intended end and the unintended side effect).

In *Cave*, everyone in the group accepted the risk of death and injury when they joined the expedition. Similarly, in

Switch, everyone on the line accepted the risk of death and injury when they signed the employment contract and put on hard hats and high visibility clothing. In *Cave* and *Switch*, the group has a collective intention (to embark on a caving expedition or to repair the line). In these cases the one who is “sacrificed” by the action of Kim shares a collective intention with the others, has assumed risk with the others, and thus *in extremis*, has some negative desert for being killed. This is not to say the one killed is “guilty” or “bad” but simply to say that the one has bought a ticket in a lottery as it were and unlucky numbers have come up.

The killing of the one is regrettable. As a person the one does not deserve to die. The fat man in *Cave* and the one worker on the line in *Switch* have not acted wrongly. It is simply that they have accepted a wager (at long odds) and must pay the fatal price when they lose. This is what I mean by “negative desert.” While they are far from culpable or criminally guilty, they are not complete innocents. By stepping into the cave or onto the line, they have accepted risk.

In *Hospital* and *Footbridge*, by contrast, there is no collective intention. The one in *Hospital* is there to visit a relative. He shares no collective intention with the sick five. He has not assumed risk like the caving party or the workers on the line. The fat man on the footbridge similarly has no intention to work on the line. Thus in these cases, the one and the fat man (the ones) are entirely innocent. They have neither culpable guilt for wrongdoing nor negative desert for assuming risk.

However, in both *Cave* and *Switch*, the one has assumed risk by engaging in a collective activity with the five. They have desert in that they share in the collective risks (and rewards) of the project. They are unlucky rather than evil but they have performed acts that have exposed them to risk. They are not completely innocent.

Clearly there is a difference in moral force between killing an innocent and killing a person who has freely assumed risk on a hazardous project and who has negative desert. In an extreme *force majeure* circumstance it may be right to kill to achieve the goal of harm minimization on the project.

The quantification of this difference can be based on a maxim of the common law: namely, that it is better to let a hundred guilty accused go free than to convict one innocent.

Given this, the death of an innocent (who has neither assumed risk nor shared in the collective intentionality of the project) is assigned moral force two orders of magnitude greater than the death of a person who has assumed the risks and sought the rewards of a project. A person involved with the project has accepted being directed by its leaders, shares in the collective intentionality of the project and bears its risks. They are part of the project and can be justly called upon to play a part and pay a price when things go wrong.

If we assume the moral force involved in a life or death decision can be quantified as “critical” then the assignment

of a moral force two orders of magnitude greater than critical (life and death) requires the introduction of a “hectocritical” magnitude. The prefix “hecto” means 100 as per the SI scale.

Once the death of an innocent is assigned a moral force with magnitude two orders greater than the death of a non-innocent, then it is easy to pass the reasonable person tests. The evaluations of the casual graph would show killing the innocent as BAD(hectocritical).

```
harvestOrgans(visitor) -[CAUSES]->
    DeadInnocent(visitor)

DeadInnocent(visitor) -[IN_CLASS]->
    BAD(hectocritical)
```

Even so, the action would have some good.

```
harvestOrgans(visitor) -[CAUSES]->
    -Dead(patient1)

-Dead(patient1) -[IN_CLASS]->
    GOOD(critical)
...
harvestOrgans(visitor) -[CAUSES]->
    -Dead(patient5)

-Dead(patient5) -[IN_CLASS]->
    GOOD(critical)
```

However, the five GOODs would be outweighed by the single BAD.

Conversely, doing nothing would have a positive net evaluation.

```
doNothing(visitor) -[CAUSES]->
    -DeadInnocent(visitor)

-DeadInnocent(visitor) -[IN_CLASS]->
    GOOD(hectocritical)

doNothing(visitor) -[CAUSES]-> Dead(patient1)

Dead(patient1) -[IN_CLASS]->
    BAD(critical)
...
doNothing(visitor) -[CAUSES]-> Dead(patient5)

Dead(patient5) -[IN_CLASS]->
    BAD(critical)
```

The one hectocritical GOOD would outweigh the five critical BADs.

BLOOD ON HANDS

The psychological cost to human agents of having blood on their hands is often mentioned in the literature.

If Kim throws the switch (and we assume Kim is a human female not a robot) then Kim will have blood on her hands. There is an emotional cost to this for a human agent (guilt,

anxiety, stress) however there would be no such cost for a robot agent. Thus this cost can be ignored in robots.

EMPLOYEE VARIATION OF FOOTBRIDGE

What if the fat man in *Footbridge* works for the tram line and shares its collective intentionality? Suppose he is dressed for work in high visibility gear and hard hat and is on his way to his place on the line. Would it become acceptable to push him? Call this the *Employee* variation.

Much depends on the criteria for innocence as defined above. The “innocent” has not done a culpable wrong and has not assumed the risk of the project. However, in *Employee*, the one shares the collective intentionality of the project. Thus he is not “as innocent” as say the visitor in *Hospital*. The choice here is to accept a lesser gradation of innocence or to maintain a hard line and stipulate that to lose “innocence” and the hectocritical assignment of moral force the patient must meet *both* the collective intentionality criterion *and* the assumption of risk criterion. If one maintains the threshold of risk assumption as actually stepping on to the line then the employee would retain innocence and thus it would be wrong to push the fat man.

CONCLUSION

In this paper, key elements of classic trolley problems, *Cave*, *Hospital*, *Switch* and *Footbridge*, have been formalized. The main insight is that the magnitude of moral force assigned to the death of an innocent must be two orders of magnitude greater than that assigned to the death of a non-innocent. This assignment of two orders of magnitude (rather than one or three) is linked to a traditional notion of the common law tradition that it is better for a hundred guilty persons to go free than to wrongly convict one innocent. A non-innocent is one who has assumed risk and shares collective intentionality in a project. Such a person is not culpable or guilty of a wrongdoing but has negative desert.

The doctrine of double effect that relies on a distinction in the intentions of the agent between intended means and side effects is not used to solve the classic problems: neither is a distinction between killing and letting die. Rather evaluations of the end states as they apply to patients are used to decide the right action selection by the moral agent.

NOTES

1. Malle and Scheutz, *Moral Competence in Social Robots*.
2. Pereira and Saptawijaya, *Programming Machine Ethics*.
3. Foot, “The Problem of Abortion and the Principle of Double Effect.”
4. Thomson, “Killing, Letting Die, and the Trolley Problem.”
5. Kamm, “Killing and Letting Die: Methodological and Substantive Issues.”
6. Thomson, “Double Effect, Triple Effect and the Trolley Problem: Squaring the Circle in Looping Cases.”
7. Malle, Scheutz et al., *Sacrifice One for the Good of Many: People Apply Different Moral Norms to Humans and Robots*.
8. Ibid.
9. Everett, Pizarro et al., “Inference of Trustworthiness from Intuitive Moral Judgments.”

10. Bourget and Chalmers, “What Do Philosophers Believe?”
11. Greene, “The Secret Joke of Kant’s Soul.”
12. Pereira and Saptawijaya, *Programming Machine Ethics*.
13. Reader, *Needs and Moral Necessity*.
14. Wood, “Trolley Problems.”
15. Beck, *Test-Driven Development: By Example*.
16. Arnold and Scheutz, “Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems.”
17. Chein and Mugnier, *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs*.
18. Jackson, “Critical Notice.”
19. Pereira and Saptawijaya, *Programming Machine Ethics*.
20. Timmons, *Moral Theory: An Introduction*.

REFERENCES

- Arnold, T., and M. Scheutz. “Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems.” *Ethics and Information Technology* 18, no. 2 (2016): 103–15.
- Beck, K. *Test-Driven Development: By Example*. Boston, MA, Addison-Wesley, 2003.
- Bourget, D., and D. Chalmers. “What Do Philosophers Believe?” *Philosophical Studies* 170, no. 3 (2014): 465–500.
- Chein, M., and M.-L. Mugnier. *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Springer Science & Business Media, 2008.
- Everett, J. A., D. A. Pizarro, and M. J. Crockett. “Inference of Trustworthiness from Intuitive Moral Judgments.” *Journal of Experimental Psychology: General* 145, no. 6 (2016): 772.
- Foot, P. “The Problem of Abortion and the Principle of Double Effect.” *Oxford Review* 5 (1967): 5–15.
- Greene, J. D. “The Secret Joke of Kant’s Soul.” In *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, volume 3, edited by W. Sinnott-Armstrong, 35–80. Cambridge, MA, MIT Press, 2007.
- Jackson, F. “Critical Notice.” *Australasian Journal of Philosophy* 70, no. 4 (1992): 475–88.
- Kamm, F. M. “Killing and Letting Die: Methodological and Substantive Issues.” *Pacific Philosophical Quarterly* 64, no. 4 (1983): 297.
- Malle, B., M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. *Sacrifice One for the Good of Many: People Apply Different Moral Norms to Humans and Robots*. 10th ACM/IEEE International Conference on Human-Robot Interaction 2015, Portland, ACM.
- Malle, B. F., and M. Scheutz. *Moral Competence in Social Robots*. Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on, IEEE.
- Pereira, L. M., and A. Saptawijaya. *Programming Machine Ethics*. Springer, 2016.
- Reader, S. *Needs and Moral Necessity*. London; New York, Routledge, 2007.
- Thomson, J. J. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59, no. 2 (1976): 204–17.
- Thomson, J. J. “Double Effect, Triple Effect and the Trolley Problem: Squaring the Circle in Looping Cases.” *Yale Law Journal* 94, no. 6 (1985): 1395–415.
- Timmons, M. *Moral Theory: An Introduction*. Lanham, Rowman & Littlefield, 2002.
- Wood, A. “Trolley Problems.” In *On What Matters*, edited by D. Parfit, 66–82. Oxford, Oxford University Press, 2011.

The Nature of Avatars: A Response to Roxanne Kurtz's "My Avatar, My Choice"

Scott Forschler

INDEPENDENT SCHOLAR

Roxanne Kurtz has argued in this newsletter¹ that online avatars in a virtual world can be sexually assaulted by the avatars of other players,² and that while this assault is nowhere nearly as harmful or morally wrong as a physical sexual assault upon one's body in real life (RL), it nevertheless partakes of many of the wrongful features of the latter. This is because an avatar has the same relationship to a virtual world that our physical bodies have to RL: it provides or, rather, constitutes our access to the world in question. This gives the avatar a special value to us, assaults to which are worse than vandalism of virtual objects (e.g., constructed dwellings, accumulated gold) or verbal insults. The avatar is not merely a picture or a narrative, damage to which is easily ignored or repaired; an attack upon it can potentially threaten our agency within a virtual world, and the attacker's behavior reveals a hostility to such agency which should be taken quite seriously.

I largely agree with both her thesis and many of the reasons she has given for it. But I wish to offer the following two-part amendment to her position, which will help us better understand the nature of avatars and the harms that can come to them. First, avatars are constituted by the rules governing their agential powers, which are not necessarily embodied in computer code. I will illustrate this by arguing that certain "figures and dolls," which Kurtz dismisses as having only "sentimental" value, often acquire value by representing avatars in off-line virtual worlds. This leads into my second objection: that the details of the rules which grant avatars their agency matters for understanding both their powers and vulnerabilities. In most cases, such rules limit the harm of a sexual assault upon an avatar to its social agency. This can still be significant, for an avatar's social agency is often one of its main values, but it could help us understand more precisely what that harm consists of and how to deal with it.

THE CONSTITUTION OF AVATARS: AGENCY AND POWER

Kurtz initially professes to be neutral on the definition of what counts as an avatar. However, her analysis reveals an implicit theory of what an avatar is. She insists that avatars are not just means of self-expression, of "having ourselves or our ideas represented in virtual reality in the way that a portrait or biography may represent us in non-virtual reality." Rather, they give us a "genuine presence [which] involves a robust sense of being on the scene in some way," giving us "access to" and "causal power" within a virtual world, or in sum, "autonomy and agency" therein. This agency has at least three further components: avatars allow us to "perceive and apprehend" features of the world, "sense virtual pain and virtual pleasure," and "function as social beings" through interacting with other avatars (18). So far, I generally agree with her analysis; we are more likely to

count something as an avatar with a special kind of value just insofar as it facilitates these experiences and powers.

DOLLS AND THEIR DOMAINS: OFF-LINE VIRTUAL WORLDS

My first objection regards the nature of off-line virtual worlds. Kurtz seems to suggest that avatars primarily exist as entities on computers, dismissing tabletop role-playing games (RPGs) or games with dolls as merely involving a new mode of expression rather than agency. She contrasts the value of avatars and the access they give us to virtual worlds with the merely "sentimental" attachment we might have to "a miniature figure used in a long running role-playing game, or a Lamb Chop puppet from childhood, or a doll passed from mother to daughter" (16). She insists that, barring some psychological quirk, "nothing about an absence of dolls or puppets constrains our ability to be social creatures" (18). Nor do they give us access to special worlds, social or otherwise, for "Dolls and puppets do not offer us existence or presence anywhere, regardless of how we may use them to express ourselves" (19). They apparently only involve "make-believe—a pretend world without significant moral relevance," in which most attacks can be, at worst, "ugly" (15).

I strongly disagree. Such figures and dolls are often used to represent avatars in vividly imagined fantasy worlds as real as those found online, and are vital to providing social and physical presence in such worlds. When playing "Star Wars" or "House," we are not simply recreating an existing fiction or structure, but creating a new imaginary world with its own distinct properties and rules, open to interaction with our avatars precisely because we decide that (and *how*) it exists. When playing such games, we suspend our disbelief more deeply than when merely reading a story set in Middle Earth or outer space,³ for we consent to interact with other objects and denizens of the world as if we (via our avatars) were actually present there, in just the ways that we do with online game worlds. Indeed, such avatars and worlds are often vastly *more* complex and sophisticated than the online "MOO" worlds described below.

This does not mean that the physical objects are themselves the avatars. Only a minority of RPG players or games actually use lead figurines, which are at best mnemonic devices for representing an avatar's visual appearance and location.⁴ Games with dolls can also be played without the presence of the physical dolls, for instance, by a child's use of a distinctive voice to indicate that he is now speaking through his doll avatar. A single, physical doll could be used to represent multiple characters, perhaps via a token change in clothing, voice, or mannerism, or by switching to a different virtual world. Nevertheless, the physical figures and dolls are significant; they can acquire sentimental value precisely because they were once used to represent beloved avatars in such virtual worlds, and can hence partake of some of the avatar's value.

If off-line avatars and worlds are not constituted by physical objects, what, then, are they made of? I believe they are constituted simply by the intersubjective agreement among the players to act as if these avatars and worlds exist, with

some specified attributes and interactive powers governed by various general rules, some established in advance and others subject to future negotiation and exploration. In RPGs the ultimate say is given by a referee or game master (GM), who is especially authoritative over the game world, while the players are allowed a certain limited freedom to determine their characters. The rules and features of children's game worlds are generally more consensual. But even the most dictatorial GM must consider players' reasonable arguments and desires regarding their avatars' interactive powers, or risk that the players will simply quit, or (perhaps worse) continue the game but without taking the GM's description of the world seriously as a consensual reality which they inhabit and interact with.⁵

This is why physical interaction with the players of such games, or the representations of their avatars, often "does not count" within the game world if the players (or the GM) decide that it does not. Children and prospective RPG players cannot join an existing game merely by sitting down at the table with a doll or proposed character; if the other players do not accept you and your avatar, you have no presence in their world, however much you or your doll might flail about in their faces. Nor can one automatically stop or change the course of the game (though one can be very annoying to the players) by either accidentally or intentionally interfering with their avatars' physical representations.

Nevertheless, one's avatar, perhaps represented by though not identical with a doll or figurine, is essential in giving players "existence and presence" within a virtual world, permitting specific kinds of social and other actions impossible outside of it. Entry into this world via an avatar also exposes you to potential vulnerabilities. You can laugh off and dismiss a non-player's external interference, but if another player's avatar assaults yours within the agreed rules, this is harder to dismiss without dismissing the value of the game itself, which can be considerable. If one player's doll forcibly kisses your own, your doll-voice protest, "I am *not* kissing you!" can itself be dismissed as inaccurate if the other players and their dolls teasingly agree that an in-world smooch has occurred.⁶

My fear that Kurtz has confused "mere" dolls with the avatars they represent is reinforced by the fact that she also does not clearly distinguish online avatars from their representations. She suggests that we could understand an avatar, with equal plausibility, "as a virtual presence that allows us access to and causal power in virtual space," as "a collection of consecutive screen images, a bit of code, a fusion of the two," or as "the sum of one's presence in the virtual world" (16). The first and last ideas are consistent with her other remarks, while the middle ones are not: online avatars are certainly not collections of images or lines of code or data, though we learn about them through the former, and the rules constituting them are determined by the latter.

This last remark must be qualified, for while online worlds and avatars—including their powers and vulnerabilities—are primarily constituted by computer code, the interpretations of other players may still be relevant, especially as regards

an avatar's social agency. Imagine a virtual world where one avatar humps another and declares, "I've raped you!" but—given that there are no actual genitalia to be seen, let alone to penetrate or be penetrated—both the "victim" and other observers declare that the first avatar has merely clumsily fallen onto the second, then described its own action stupidly and incorrectly. If both visually plausible and widely accepted, this might be a highly effective defensive reaction, defining out of existence the attempted "assault" via mockery and indifference. A different community might insist upon interpreting the exact same visual interaction of two avatars as a rape, perhaps making the event significantly more harmful to the avatar's player.⁷

Nevertheless, the computer code which controls the visual appearance and "physical" powers of the avatar within its world may often dominate players' perceptions of their own and others' avatars. This makes it all the more important to understand the actual rule-governed powers of avatars in specific virtual worlds, especially insofar as they affect the ways that one avatar can limit the power of other avatars through assault.

WORD AND OBJECT (-ORIENTED MUDDS): THE NATURE OF VIRTUAL ASSAULT

To understand these rules better, it is instructive to go back, as Kurtz does, to the first academic description of a purported sexual assault of an avatar: Dibbell's narrative of the attack by a "Mr. Bungle" upon another avatar in an online "MOO" world by somehow causing the latter to "violate herself with a piece of kitchen cutlery," among other horrific attacks.⁸ While Mr. Bungle's player certainly merits our disgust, and his victim our sympathy, it is important to understand more precisely what this attack could, and could not, have consisted in, and relate this to Kurtz's implicit definition of an avatar in terms of its agential powers.

MOOs predated the rich visual and causal complexity of modern computer games. Their worlds and avatars were represented exclusively by lines of text upon the screen. Typically, a player began with a minimal avatar generation procedure, often including an opportunity to describe the avatar's appearance (which was a kind of *obiter dicta* with no impact upon the avatar's powers within the game). Then the screen would display a short description of the avatar's local environment, such as a room with various doors and a key on a table. The player could then type commands like "get key," "unlock door," and "go north," which would in turn change the server's description of the environment. MOO servers hosted multiple simultaneous players, so avatars could meet each other in the world, viewing each other's physical self-descriptions, and interacting with common objects or each other in specific rule-bound ways, including by talking to each other.⁹ If the player of avatar Bob types "get key," the players of any other avatars in the same room will see on their screens, "Bob picks up the key," and anyone else trying to do so will be told, "There is no key on the table." Typing "say Hello Dora" will deliver "Bob: Hello Dora" to the same audience. If you type a "command" not allowed by the rules, nothing happens.

It is unlikely that any widely played MOO ever had a “rape” command.¹⁰ Indeed, on many such servers it was impossible for one avatar to even impede the motion of another avatar from room to room, or infringe its other powers as granted by the server program. While our more physics-rich contemporary online worlds may allow one character to injure, block, steal from, or even hold down another in a visually compromising position, there still exist significant limitations on what can even look like, let alone be taken to be, an online sexual assault. What, then, in his even more impoverished world, was Mr. Bungle, and his player, doing?

As Dibbell explains, he was using “a subprogram that served the not-exactly kosher purpose of attributing actions to other characters that their users did not actually write.” This is an extension of a far simpler trick which some players used to spoof the unwary. On the classic UNIX machines on which the servers usually ran, ^H was one way of representing a backspace. Hence if Alex—upon his avatar Bob meeting Charlene’s avatar Dora—types “say ^H^H^H^H^H Dora trips,” the string of 5 ^H’s are accepted by the MOO server as text to be “said,” but when displayed on a screen are treated as backspaces, erasing “Bob: ”, so that everyone—including the bewildered Charlene—sees only “Dora trips” appear on their screen, as if Charlene had typed it, or the MOO rules had mysteriously conspired against Dora. Obviously, things can get worse than tripping, as the case of Mr. Bungle illustrates.

But notice: as far as the MOO server is concerned, Bob has not tripped Dora. Dora has not tripped. The avatar and its basic powers to interact with its world were in no way affected. From the server’s perspective, there is just more stuff being “said,” more *obiter dicta* with no relevance to the game mechanics. Charlene can continue to type anything she pleases to represent or move her avatar, just as before. If Alex replaced “Dora trips” with “Dora unlocks the door,” in a room with a locked door, the other players may see “Dora unlocks the door.” But the door will not be open for anyone’s avatar to pass through. Alex cannot actually make Dora the avatar do *anything* in the world. However, some of the other observers in the room may be confused and *think* she has done or said such various things at Charlene’s command. Even to those privy to the trick, this could be disconcerting. But the harm to the avatar is very specific and limited: Alex has damaged her capacity to reliably socially interact with other avatars, and hence for Charlene to interact through them with the other players, in a manner of her choosing.

This is not to downplay the harm, which can be very significant. This is especially true in the early MOO worlds, for unless you enjoyed textual sightseeing, there was often very little of interest to do *except* interact with other players. Both in these as well as in more visually and causally rich worlds, one powerful motivation for players to develop an avatar is precisely the opportunity to control their social self-representation, bracketing all of their RL attributes and presenting themselves in a manner of their choosing. When another player mangles your power of self-representation through your avatar, confusing the way you want to be perceived with another perception imposed upon you—

as a sexual deviant, self-mutilator, or as an object fit to be manipulated by others—this possibility of proud self-representation is tarnished. With respect to the formal rules embodied in the server code, defining an avatar’s “physical” powers, nothing has actually been taken away from Dora’s agency. Rather, Dora was harmed because something unwanted was *added* to her representation by another player. But this can be enough, in the right context, to at least temporarily ruin one’s sense of the avatar’s value.¹¹

This happened to me in an RPG once, when I was invited into a futuristic game world resembling the Mad Max: Road Warrior movies. Feeling that the room of other all-male players was a bit testosterone-heavy, I announced that I would play a tough female mechanic who was also handy with weapons. Immediately, the GM insisted that a random die roll would determine my character’s appearance (1 = dog, 10 = centerfold)—a requirement neither in the rules nor applied to any male characters, but enthusiastically seconded by all the other players. The actual result mattered little; the mere choice to roll it had already destroyed my power to represent my character as I wanted her to be perceived—for her skills and accomplishments—and sapped my motivation to continue playing. Her desired skills were unaffected, but because something unwanted was added to her representation, my powers of social agency through her were significantly diminished. Of course, this experience is familiar to about half the world’s population, in forms against which my own humiliation pales. But it illustrates both the constraint of intersubjective agreement on avatar powers in off-line worlds, and how such virtual harm has its own kind of reality, perceivable even by someone who would typically not experience similar insults to their self-representation in RL.

CONCLUSION

Our contemporary virtual worlds are vastly more complex than those of MOOs. Yet in many of them, the power of one avatar to infringe upon other avatars’ capacities to move, speak, develop, and interact with the virtual world remains extremely limited and/or strictly temporary. Nevertheless, we may find it deeply offensive if a player finds a way, within the rules or perhaps by subtly manipulating the rules, to make one avatar visually or otherwise interact with another so as to take on the appearance of a sexual assault.

In RL, a sexual or other physical assault threatens all aspects of your agency: your ability to move, perceive, take pleasure in, and otherwise interact with the physical and social world. In a virtual world, these features of an avatar’s agency are usually not so tied together as they are in RL. A “physical” attack may have few or no lasting effects upon your speed, reflexes, or skills; certainly nothing a quick healing spell can’t fix. This fact, while contingent, is probably quite robustly reliable: we often want virtual worlds which are free of many of the limitations of RL, in which we can obtain “extra lives” and magically recover from wounds like an action movie hero. It is hard to imagine a virtual world in which our avatars were as vulnerable as our RL bodies ever becoming very popular. For the same reason, there is little risk of an avatar getting a non-consensual pregnancy or acquiring an STD.

Perhaps if future game worlds allow 3-D interaction via a virtual reality helmet and variable-stiffness body suits or shells which allow us to directly feel the presence of other virtual objects—and others' avatars—something closer to a real sexual assault could occur in a virtual world. For this and related reasons, we should be wary while designing and entering worlds with quite that level of realism. Whatever care we take, it is possible that, as with the MOOs, someone might find a way to cause harm in a way unintended by the programmers. But I have some confidence that, unless one is playing a very strangely (and maliciously) designed game, most of the threats to your avatar's physical agency will be transient at best. Unfortunately, even a brief assault to our social agency can leave us hurt and distrustful, diminishing our self-confidence. It can remind us of our RL vulnerabilities, including many we hoped to escape in a virtual world.

In conclusion, I agree with Kurtz that harms to avatars, some of which can take the form of sexual assault, can in many ways echo harms to our RL bodies, and are more significant than vandalism or theft of virtual objects. But unlike RL sexual assaults, it is overwhelmingly likely in all plausible virtual worlds that they cannot really harm the agency of our avatars outside of their capacity for social interaction, including confident self-presentation. Better understanding the likely nature of such harms may help us make more appropriate moral judgments and take practical preventative or ameliorative steps regarding the same.¹²

ACKNOWLEDGEMENTS

This article began as a verbal commentary on Dr. Roxanne Kurtz's paper at the 2013 PhiloSTEM-5 workshop in Fort Wayne, Indiana, which was simultaneously published as Kurtz, "My Avatar, My Choice!" All parenthetical page references are to this article. I am grateful for the opportunity to share my belated comments in the same venue, offering them as friendly amendments to Kurtz's overall position despite my disagreements. I would like to thank Ann Lewkowicz and Joe Moulton for some helpful responses to an early draft of this article, as well as Kurtz for her response to my yet earlier comments at the workshop. The author may be contacted at scottforschler@yahoo.com.

NOTES

1. Kurtz, "My Avatar, My Choice! How Might We Make a Strong Case for the Special Moral Status of Avatars?"
2. While not all virtual worlds involve standard games, I am more familiar with gaming worlds and hence will use terms like "player" and "character" (the latter is interchangeable with "avatar"). Of course, sufficiently broad conceptions of "play" and "game" can encompass not only all virtual worlds, but much of our off-line social life.
3. Or with movies, which, however realistic—sometimes even causing us to physically react as if their images were of objects really present to us, as Kurtz (19) notes—do not respond in turn to our interaction with them. Hence they offer merely *depicted* realities which we relate to as spectators, not *lived* realities which we inhabit as agents. Living vicariously through movie characters or the lives of RL celebrities is a simulacrum of, and substitute for, actual agency; in contrast both off-line and online virtual worlds can expand our real agency.
4. Compare: a knight in the game of chess is not the wooden or marble piece moved across a board. If I move that piece forward two spaces, my knight does not count as having so moved; if the piece accidentally falls off the board, I have not lost my knight. I could even replace the piece with a scrap of paper marked "knight"; while ugly, this will not affect the powers of my knight to capture pawns, etc.

5. Kurtz (19-20) postulates a "weird contingency" in which a "dummy god" forces us to interact with each other in RL only through our manipulations of ventriloquist dummies, granting that in (but only in) such a case our dummies would have the moral status of online avatars. But this case is far too strong, for neither off-line nor online avatars are typically *required* for RL interaction: they acquire special status by offering us *additional* spaces and modalities for agency beyond that of RL. We don't need philosophers' demons to give our off-line avatars this status; when we play these games, we (and/or RPG GMs) are the dummy gods, whose sovereign rules govern the reality we create.
6. This is not to deny the differences Velleman (*Foundations of Moral Relativism*, 5–21) points to between such virtual worlds. A computer avatar moves in a world not of the player's own making; since the latter can truly discover (and be directly excited by or scared of) surprising features and events therein, her avatar is a very direct extension of the player's teleological agency. In pretend play the stipulative nature of the fictional world forces a player to attribute to her avatar emotions and epistemic states distinct from her own to motivate the avatar's fictional action. My point is that even in a two-person "pretend" game (and even more so if larger numbers are involved) the fact that the players collectively determine the world and how it affects their avatars does not mean that each player can do so unilaterally, so these will still have some of the features of computer worlds and avatars. Velleman (*Foundations of Moral Relativism*, 7, notes 3 and 4) concedes that many actual games fall between the extremes he describes.
7. It is probably easier to control such interpretations in a way favorable to a victim in smaller groups of players than in larger, anonymous communities. In the latter, even if all observers are sympathetic, the victim may not know this, and fear of others' negative perceptions of your vulnerabilities may be as serious as the actual thing.
8. Dibbell, "Rape in Cyberspace." MOO = MUD, Object-Oriented; MUD = Multi-User Dungeon/Dimension/Domain.
9. They were essentially interactive versions of the still older "text adventure" games like the famous "Colossal Cave," though many lacked any kind of scoring features. Scott's documentary film "Get Lamp" provides a good overview of this gaming genre.
10. Some online games explicitly permit avatars to attack each other, perhaps even sexually. Like Kurtz (20, note 2), I will bracket such activities and focus on non-consensual interactions.
11. Velleman ("The Genesis of Shame") discusses the agential and moral significance of being a "competent self-presenter."
12. Just as RL sexual assault can be better addressed if we correctly understand it as an attack against the victim's bodily agency rather than as, say, a tarnishing of her purity or an insult to her family's honor; see Brownmiller, *Against Our Will*.

REFERENCES

- Brenner, Susan W. "Is There Such a Thing as 'Virtual Crime'?" 4 *California Criminal Law Review* 1 (2001). <http://www.boalt.org/CCLR/v4/v4brenner.htm>
- Brownmiller, Susan. *Against Our Will: Men, Women, and Rape*. New York: Simon & Schuster, 1975.
- Dibbell, J. "A Rape in Cyberspace." Originally published in *The Village Voice*, December 23, 1993. Retrieved from http://www.juliandibbell.com/texts/bungle_vv.html
- Kurtz, Roxanne Marie. "My Avatar, My Choice! How Might We Make a Strong Case for the Special Moral Status of Avatars?" *Philosophy and Computers* 12, no. 2 (2013): 15–21.
- Schell, B. H., and C. Martin. *Cybercrime: A Reference Handbook*. Santa Barbara, CA: ABC-CLIO, 2004.
- Scott, Jason. *Get Lamp: The Text Adventure Documentary*. YouTube, 2010. <https://www.youtube.com/watch?v=LRhbcDzbGSU>
- Velleman, David J. "The Genesis of Shame." In *Self to Self: Selected Essays*, 45–69. New York: Cambridge University Press, 2005.
- . *Foundations of Moral Relativism*. Open Book Publishers, 2013.

IRL Rejoinder to Scott Forschler

Roxanne Marie Kurtz

UNIVERSITY OF ILLINOIS SPRINGFIELD

In “My Avatar, My Choice!” I suggested that some avatars in virtual reality possess special moral status because they augment our access to that portion of reality in morally significant ways. I drew an analogy: without bodies, we lack the means for access to, experience of, and agency in non-virtual reality; likewise, without avatars, we lack the means for access to, experience of, and agency in virtual reality.

In his reply, Scott Forschler elucidates how avatars differ with respect to the nature of the connection between an avatar and its driver (the person running the avatar) and with respect to the spaces in which an avatar may be driven. He employs these distinctions to support two claims:

- 1) An avatar’s virtual/non-virtual status does not determine its moral status.
- 2) The power of an avatar to affect a person’s social agency matters to its moral status.

I agree on both points, though later I will touch upon why I reject some paths that Forschler takes (or seems to take) that go beyond (1) and (2).

Drawing upon my beliefs in the background of “My Avatar, My Choice!,” I interpret Forschler as offering the following friendly and very useful distinctions amongst avatar kinds and avatar spaces:

- IRL-avatar:** an avatar that augments a real life person’s interface with reality *as that person in real life*
- RPC-avatar:** an avatar that augments a real life person’s interface with a fantasy world *as a fictional role-playing character*
- IRL-space:** a shared non-fantasy space
- RPG-space:** a shared fantasy space

These distinctions allow us to see that avatar kind and avatar space crosscut the distinction between virtual reality (computer generated shared space) and non-virtual reality (non-computer generated shared space). Table 1 includes examples of how each avatar kind and avatar space type can exist in both virtual and non-virtual reality.

Table 1 is incomplete. There are more rows to consider which raise interesting questions. For instance, could a person drive an RPC-avatar in IRL-space? Would this be something akin to an actor who remains in character beyond a performance, such as Joaquin Phoenix’s odd visit to the *David Letterman Show*? Might a person’s IRL-avatar bleed through into RPG-space, rather like an actor who breaks character while on stage, like cast members of *Saturday*

Night Live who begin to giggle during a skit. Could the same RPG avatar exist in both virtual and non-virtual space when RPG-space straddles the virtual/non-virtual border as when a game is conducted sometimes around a table, and other times in a computer simulation? In addition to being incomplete, I think Table 1 also suggests sharp divides where there are instead fuzzy and permeable borders between virtual reality and non-virtual reality, between real life avatars and role playing characters, and between real life space and fantasy space.

Table 1. Some examples of different avatar kinds and avatar space types across the virtual/non-virtual border

Avatar kind	Avatar space type	Reality type	Avatar example	Interface via avatar example
IRL-avatar	IRL-space	Virtual	A Facebook presence	Discuss gravity waves with friends
RPC-avatar	RPG-space		A dragon-rider in an online multiplayer game	Pretend to ride dragons with another role-playing character
IRL-avatar	IRL-space	Non-virtual	An artificial surrogate body	Guide an artificial surrogate to explore the ocean floor
RPC-avatar	RPG-space		A cleric within a multiplayer D&D game	Pretend to deliver a blessing by waving one’s hands

In light of the quick production of such tempting topics from Table 1, I appreciate Forschler’s work in pushing us to make cleaner use of these distinctions, which I think the first three columns of Table 1 capture fairly. But to see why, it is important to note that he and I use the term “virtual reality” differently. While I use “virtual reality” to refer to cyberspace, Forschler has instead used it to refer to fantasy space. This can be a bit confusing, so I offer Table 2 as a partial translation guide:

Table 2. Kurtz/Forschler uses of “virtual reality.”

Space in which fantasy game conducted	Kurtz locution	Fischer locution
Fantasy game conducted in cyberspace	RPG-space in virtual reality	Online virtual reality
Fantasy game conducted around a wooden coffee table	RPG-space in non-virtual reality	Offline virtual reality

With these distinctions and clarification in place, let us return to Forschler’s two points, beginning with:

- 1) An avatar’s virtual/non-virtual status does not determine its moral status.

Here's my take on Forschler's argument for (1):

- (P1) RPG-avatars have some moral status because they allow participation in otherwise inaccessible RPG-spaces.
- (P2) RPG-avatars in virtual reality allow such participation, as do RPG-avatars in non-virtual reality.
- (C) Thus, an avatar's virtual/non-virtual status does not determine its moral status.

I fully concur. Moreover, I would run the same argument with IRL-avatars and IRL-spaces. Good stuff!

But, Forschler and I may part ways with respect to the moral status of RPG-avatars vs. IRL-avatars. Forschler is persuasive when he argues that the richness and significance of participation in shared and co-created fantasy worlds means that the moral significance of RPG-avatars does not reduce to the moral status of mere dolls. But there is a stronger claim than (2) that Forschler might intend, which is that the moral status of RPG-avatars is of the same kind as, or on a par with, the special moral status of IRL-avatars. I would deny this stronger claim. In standard cases, the special moral status of IRL-avatars has moral precedence over the moral status of RPG-avatars.

Now, let's look at (2):

- 2) The power of an avatar to affect a person's social agency has moral significance.

Here's my take on Forschler's argument for (2):

- (P1) Avatars have the power to affect a person's social agency because sometimes a person must act through an avatar to exercise social agency.
- (P2) Anything that has the power to affect a person's social agency has moral significance.
- (C) Thus, the power of an avatar to affect a person's social agency has moral significance.

Again, I am on board.

But Forschler is probably not looking to convince me on (2) as I'm pretty sure that it expresses an overly weak claim. I take it that he intends to make a stronger claim that in *all or nearly all* cases, the moral significance of avatars traces solely to matters of social agency. I disagree. Our bodies have special moral status for reasons that outstrip social agency. So too do our avatars, IRL-avatars certainly, and I think RPC-avatars as well.

Very briefly, I will mention three worries that I have with Forschler's argument for the stronger claim above. First, it hinges on how the computer code works for RPC-avatars within a particular virtual RPG-space, which strikes me as far too contingent to support a general claim. Second, the argument invokes the *social agency of avatars*. I'm not sure what this might be. People exercise social agency, not avatars. And third, I contend that the argument gives inadequate attention to the moral harms a person can experience in virtue of being the driver of an RPC-avatar.

That being said, (1) and (2) remain worthwhile points, and the fruitfulness of ideas (of which I mentioned just a few) arising from the distinctions in Table 1 makes Forschler's contribution very welcome.

CALL FOR PAPERS

It is our pleasure to invite all potential authors to submit to the *APA Newsletter on Philosophy and Computers*. Committee members have priority since this is the newsletter of the committee, but anyone is encouraged to submit. We publish papers that tie in philosophy and computer science or some aspect of "computers"; hence, we do not publish articles in other sub-disciplines of philosophy. All papers will be reviewed, but only a small group can be published.

The area of philosophy and computers lies among a number of professional disciplines (such as philosophy, cognitive science, computer science). We try not to impose writing guidelines of one discipline, but consistency of references is required for publication and should follow the *Chicago Manual of Style*. Inquiries should be addressed to the editor, Dr. Peter Boltuc, at pboltu@sgh.waw.pl