APA Newsletters

NEWSLETTER ON PHILOSOPHY AND COMPUTERS

Volume 12, Number 2 Spring 2013

apa
The American Philosophical Association
University of Delaware
Newark, DE 19716

 ISSN 2155-9708

APA Newsletter on

Philosophy and Computers

Piotr Bołtuć, Editor | Spring 2013 | Volume 12, Number 2

From the Editor

The first two works in the current issue of the newsletter discuss consciousness and its phenomenal nature. We are delighted to publish a version of the paper by Terry Horgan that was presented at the session organized by this committee at the APA Central Division Meeting in the spring of 2012. Horgan discusses the issue of mental intentionality and its impact for machine consciousness. The author argues that true understanding requires mental intentionality. Horgan closes with an interesting transformation of Searle's Chinese Room scenario going through various simulations, which helps clarify intuitions. This article is followed by Riccardo Manzotti's cartoon pertaining to the very same issue.[1]

This is followed by papers that discuss the ontological and moral status of avatars. Both Roxanne Kurtz and Sidey Myoo (Michał Ostrowicki) defend a special status of online beings. Kurtz focuses on ontological arguments, while Myoo also talks about practical applications, such as the PhD and MA defenses in philosophy that took place in Second Life in Poland. These are followed by Ron Loui's commentary on a recent book by Frederick Schauer, concerning the issue of defeasibility.

The last two papers are reports of work in progress. Colin Allen and his team present new results on philosophical mapping (much earlier outcomes were published in this newsletter in the past). Last, but not least, I am pleased to publish a contribution by Federico Gobbo that presents the singularity debate with a critical eye. This is a great paper to start further debate on this philosophically controversial topic.

Peter Boltuc

Note

1. Let me mention, as I do in my introduction to Manzotti's cartoon published in this newsletter almost every year, philosophical cartoons are a bit deceitful since they are more persuasive yet less argumentative than regular papers.

Articles

The Real Moral of the Chinese Room: Understanding Requires Understanding Phenomenology

Terry Horgan
University of Arizona

I have three main goals in this paper. First, I will briefly summarize a number of claims about mental intentionality that I have been articulating and defending in recent work (often collaborative, with George Graham and/or Uriah Kriegel and/or John Tienson).[1] My collaborators and I contend that the fundamental kind of mental intentionality is phenomenal intentionality. Second, I will set forth some apparent implications of this conception of mental intentionality for philosophical issues about machine consciousness—and, specifically, implications concerning the status of John Searle's famous "Chinese Room" argument. The real moral of the Chinese Room, I maintain, is that genuine understanding requires understanding phenomenology—a species of so-called "cognitive phenomenology." Third, I will give a thought-experimental argument for the existence of language-understanding cognitive phenomenology. The argument will commence from Searle's Chinese Room scenario, will proceed through a sequence of successive variations on the scenario, and will culminate in a clearly conceivable scenario that makes manifest how different a real language-understander would be from someone who lacked any language-understanding phenomenology.[2]

Phenomenal intentionality

The most basic kind of mental intentionality, according to my collaborators and me, is phenomenally constituted, and is narrow; we call it *phenomenal intentionality*. It is shared in common with all one's metaphysically possible phenomenal duplicates, including one's brain-in-vat phenomenal duplicate and one's Twin Earth phenomenal duplicate. Aspects of phenomenal intentionality include the (distinctive, proprietary, individuative) "what it's like" of (1) sensory-perceptual phenomenology, (2) agentive phenomenology, and (3) propositional-attitude phenomenology (including (3a) the phenomenal character of *attitude-type* (e.g., belief-that, wondering-whether, wanting-that, etc.), and (3b) the phenomenal character of *content* (e.g., *that Obama was reelected*, *that-Karl Rove was furious about Obama's reelection*, etc.) Some kinds of mental *reference* (e.g., to shape-properties and relative-position properties) are secured by experiential acquaintance with apparent instantiations of these properties in one's apparent ambient environment. (Such mental reference

is shared in common with one's brain-in-vat phenomenal duplicate.) Other kinds of mental reference (e.g., to concrete individuals like Karl Rove, and to natural kinds like water) are secured by the interaction of (a) phenomenal intentionality, and (b) certain externalistic connections to actual individuals or properties in one's ambient environment. (One's brain-in-vat phenomenal duplicate suffers mental *reference failure* for its thought-constituents that work like this, e.g., its *Karl Rove* thought-constituent and its *water* thought-constituent; one's Twin Earth phenomenal duplicate refers to Twin-Karl with its *Karl Rove* thought-constituent, and to XYZ with its *water* thought-constiuent.) Contrary to Quine's influential arguments for the indeterminacy of content in language and thought, phenomenal intentionality has determinate content—which in turn grounds content determinacy in public language. (What it's like to think "Lo, a rabbit" is different from what it's like to think "Lo, a collection of undetached rabbit parts.")

Implications for machine consciousness, and for the status of Searle's "Chinese room" argument

Assuming that the above claims are correct, what are the implications concerning machine consciousness and machine understanding? Well, in order for a machine to have genuine consciousness and understanding—including full-fledged, conscious, underived, content-determinate, mental *intentionality*—it would need to have *phenomenology* of a kind that includes phenomenal intentionality. More specifically, in order for a machine to have genuine language-understanding, it would need to have *language-understanding* phenomenology—a species of cognitive phenomenology.

In light of this, consider Searle's Chinese Room scenario. The guy in the room certainly has no language-understanding phenomenology, and thus doesn't understand Chinese. Also, the whole room setup, with the guy in the room as a functional component, certainly has no language-understanding phenomenology either, and thus doesn't understand Chinese. So Searle was right to claim that a machine couldn't understand Chinese just in virtue of implementing some computer program. And this conclusion generalizes the following: a machine couldn't understand Chinese just in virtue of implementing some specific form of functional organization—whether or not that functional organization is specifiable as a computer program.[3] The trouble is that the functional, causal-role features of the internal states of a machine (or a human) are entirely relational and non-intrinsic—whereas phenomenal character is an intrinsic feature of mental states, qua mental.

What then would be required in order for a machine to have genuine language-understanding? Well, the machine would need to have language-understanding phenomenology—something that would be intrinsic to its pertinent internal states, qua mental, and which therefore would not consist merely in the causal-functional roles played by those internal states. In order to build a machine that really understands language, therefore, one would need to build into the machine whatever feature(s) constitute a nomically sufficient *supervenience base* for Chinese-understanding phenomenology.

What might such a supervenience base consist in? One conjecture is that some specific form of functional architecture, when operative, constitutes a nomically sufficient supervenience base for intrinsic language-understanding phenomenology—even though genuine understanding itself consists not in the non-intrinsic, purely relational, causal-functional roles played by the physical states that implement the operation of the functional architecture, but rather in the supervenient phenomenology. Presumably, then, it would be possible in principle to engineer certain machines or robots, with control systems built out of wires and silicon chips and the like (rather than biological neurons), that possess genuine understanding, including genuine language-understanding. But Searle would still be right: genuine understanding would be present not in virtue of the non-intrinsic, causal-relational features of the implementing physical states, but rather by the supervenient phenomenology—which is intrinsic qua mental. (There is a looming worry, though, that for any proposed functional architecture, it will always be possible to invent some screwball form of implementation—along the lines of Searle's Chinese Room—that leaves out the phenomenology.)

A more plausible conjecture, I suggest, is that the needed kind of supervenience base would have to be not just some specific kind of operative functional architecture, but rather some specific kind of *implementation* of some suitable functional architecture. A serious possibility, I think, is that the right kind of physical implementation could be characterized fairly abstractly, while yet still describing certain physically intrinsic aspects of the implementational states rather than mere causal-relational aspects. A further serious possibility is that such abstractly described intrinsic physical features of the requisite implementational states would be *physically multiply realizable*—and, moreover, would be physically realizable not only within brains composed of biological neurons but also within suitably engineered machines or robots whose control circuitry is composed of the kinds of hardware found in computers. (The idea is that an abstractly described intrinsic feature of physical states could be realized by various different kinds of concrete physical states—much as a given temperature-state of a gas can be physically realized by numerous different concrete configurations of the constituent gas-molecules.) But once again, Searle would still be right. Real mentality in these machines—including real mental intentionality in general, and real Chinese-understanding in particular—would obtain not in virtue of operative functional architecture, and not in virtue of some specific mode of physical realization of that functional architecture, but rather in virtue of the understanding phenomenology that *supervenes* on that architecture as so realized—an intrinsic aspect of understanding states qua mental.[4] (The "hard problem" of phenomenal consciousness would now include the question of why such-and-such abstractly described physical feature of an implementing state—a feature of the state that is intrinsic *qua physical* (albeit also abstract and multiply realizable)—should be accompanied by so-and-so phenomenal character—a feature that is intrinsic *qua mental*.)

From the Chinese room to cognitive phenomenology: a morph-sequence argument

The claims and conjectures I advanced in the previous section presuppose the general conception of mental intentionality sketched in section 1. In particular, they presuppose the existence of (distinctive, proprietary, individuative) *cognitive phenomenology*—and, specifically, language-understanding phenomenology. But there is currently an active debate in philosophy of mind about whether there is such a thing as cognitive phenomenology. Most parties to this debate agree that there is such a thing as phenomenal consciousness, and that it includes sensory-perceptual phenomenology. Many who profess skepticism about cognitive phenomenology also acknowledge that sensory-perceptual phenomenal states are inherently intentional. And many of the skeptics acknowledge one or another kind of phenomenology other than sensory-perceptual—e.g., sensory-imagistic phenomenology and/or emotional phenomenology. But the skeptics deny the existence of *cognitive* phenomenology—viz., distinctive, proprietary, and individuative phenomenology inherent to occurrent, conscious,

propositional-attitude states. The skeptics would also deny that there is any distinctive, proprietary, and individuative phenomenology of occurent understanding-states, such as the state of understanding a just-heard Chinese utterance.

Arguing in favor of cognitive phenomenology is a tricky business. After all, phenomenological inquiry is primarily a first-person, introspective process—and the skeptics claim that when they themselves introspectively attend to their own experience, they can find no cognitive phenomenology. Dialectical progress is still possible, though. One useful approach is what is sometimes called the method of *phenomenal contrast*: describe two kinds of experience that all parties to the debate can agree are both conceivable and are distinct from one another; then argue, abductively rather than by direct appeal to introspection, that the best explanation of the difference is that one experience includes the disputed kind of phenomenology, whereas the other experience does not.[5]

I propose now to offer a new argument in favor of cognitive phenomenology—and, more specifically, in favor of the (distinctive, proprietary, individuative) phenomenology of language-understanding.[6] The argument will deploy the method of phenomenal contrast, and will proceed step-wise through a "morph" sequence of thought-experimental scenarios, each being a coherently conceivable scenario involving a guy who does not understand Chinese. Regarding early stages in the sequence, skeptics about cognitive phenomenology may well think that the guy's lack of understanding is readily explainable without positing proprietary language-understanding phenomenology. By the end of the sequence, however, the only credible potential explanation for the guy's inability to understand Chinese will be that he lacks Chinese-understanding phenomenology.

Stage 1: Searle's famous Chinese Room thought experiment. One can intelligibly conceive the guy in the room, following symbol manipulation rules in the way Searle describes. The guy in the room understands no Chinese at all; surely everyone would agree about that. And that is all I need, for present purposes.

Stage 2: The guy is still in the room. But the manipulation of the symbols that come into the room is done not by the guy himself, but (very rapidly) by a monitoring/processing/stimulation device (MPS device) appended to the guy's brain. The MPS device monitors the visual input coming into the guy's eyes, takes note of the input symbols (in Chinese) the guy sees, rapidly and automatically executes the symbol-manipulation rules, and then stimulates the guy's brain in a way that produces totally spontaneous decisions (or seeming-decisions) to put certain (Chinese) symbols into a box. Unbeknownst to the guy, the box transmits these symbols to the outside world, and they are answers in Chinese to questions in Chinese that were seen by the guy and manipulated by the MPS device. The guy in the room understands no Chinese at all; surely everyone would agree about that.

Stage 3: The Chinese-language questions now come into the room in auditory form; they are heard by the guy, whose auditory inputs are monitored by the MPS device. The MPS device rapidly and automatically executes the symbol-manipulation rules (rules that take auditory patterns as inputs), and then stimulates the guy's brain in a way that produces totally spontaneous decisions (or seeming-decisions) to make various meaningless-to-him vocal noises. Unbeknownst to the guy, the meaningless-to-him sounds he hears are Chinese-language questions, and the meaningless-to-him vocal noises he finds himself spontaneously "deciding" to produce are meaningless-to-him Chinese-language answers that are heard by those in the outside world who are posing the questions. The guy in the room understands no Chinese at all; surely everyone will agree about that.

Stage 4: The Chinese-language questions again come into the room in auditory form; they are heard by the guy and are monitored by the MPS device. The guy now sees out of the room, through a scope; he sees the people who are producing the Chinese-language questions, and he also sees and hears others who are conversing with one another while engaging in various forms of behavior (including the use of written Chinese script). But the guy also has a serious memory deficit: he persistently lacks any memories (either episodic or declarative) that extend further back in time than thirty seconds prior to the current moment. Because of this, he is unable to learn any Chinese on the basis of what he sees and hears. The MPS device rapidly and automatically executes the symbol-manipulation rules (applying them to the auditory and visual inputs emanating from those people outside the room who are looking straight toward the guy), and then stimulates the guy's brain in a way that produces totally spontaneous decisions (or seeming-decisions) to make various meaningless-to-him vocal noises in response to the meaningless-to-him sounds that he hears coming from those people outside the room who he sees are looking directly toward him when making those sounds. The guy in the room understands no Chinese at all, and cannot learn any Chinese because of his memory deficit.

Stage 5: Several aspects of stage 4 get modified at this stage. The modifications are substantial enough that it will be useful to sort them into four separate sets of features, as follows.

(1) The MPS device now monitors all the guy's sensory inputs (not just visual or auditory inputs). It also monitors all his occurrent desires and beliefs and other mental states (both present and past). It constantly stimulates his brain in ways that generate spontaneous decisions (or seeming-decisions) on his part to move his body in ways that are suitable to the overall combination of (a) the guy's beliefs and desires and other mental states (both present and past, many of which are, of course, currently forgotten by the guy himself) and (b) the content of his current sensory input (including the content of the meaningless-to-him sign-designs that happen to be written Chinese or spoken Chinese).

(2) The MPS device generates in the guy any (non-cognitive) sensory images, (non-cognitive) emotional responses, and other non-cognitive phenomenology that would arise in a guy who (a) understood Chinese, (b) had normal memory, and (c) was mentally and behaviorally just like our guy.

(3) The MPS device prevents from occurring, in the guy, any conscious mental states that would normally, in an ordinary person, accompany mental states with features (1)–(2) (e.g., confusion, puzzlement, curiosity as to what's going on, etc.). This includes precluding any non-cognitive phenomenology that might attach to such states.

(4) Rather than being stuck in a room, the guy is out among the Chinese population, interacting with them both verbally and nonverbally. He is perceived by others as being a full-fledged, ordinary, understander of Chinese.

This guy understands no Chinese at all.

Each of these successive stages is coherently conceivable, I submit. And for each scenario, it seems obvious that the guy understands no Chinese. One might well wonder about the MPS device, especially as the stages progress. Might the device

have full-fledged mental intentionality at some stage in the sequence? Might it understand Chinese? Perhaps or perhaps not, but it doesn't matter for present purposes. The key thing is that *the guy himself* understands no Chinese; the MPS is external to the guy's mind, even if it happens to have a mind of its own.

Scenario 5 is the one I now want to focus on, harnessing it for use in an explicit argument by phenomenal contrast. There is a clear mental difference between this guy (as I'll now continue to call him) and another guy we might envision (who I'll call "the other guy"). The other guy is someone who goes through all the same social-environmental situations as this guy and exhibits all the same externally observable behavior, who has ordinary memory, who understands Chinese, and whose mental life is otherwise just like this guy's.

Now comes the key question: What *explains* the mental differences between this guy and the other guy? The only adequate explanation, I submit—and therefore the correct explanation—is the following: *this guy lacks Chinese language-understanding phenomenology* (and also lacks memory-phenomenology), whereas the other guy (who is psychologically normal) undergoes such phenomenology. Hence, by inference to the best explanation, ordinary human experience includes language-understanding phenomenology (and also memory phenomenology).

Skeptics about cognitive phenomenology typically try to resist arguments from phenomenal contrast by saying that the contrasting scenarios can be explained in terms of mental differences other than the presence versus absence of cognitive phenomenology. Consider, for instance, the case of two people side-by-side both hearing the same spoken Chinese, one of whom understands Chinese and the other of whom does not. Advocates of cognitive phenomenology like to point to such cases, claiming that there is an obvious phenomenological difference between the two people even though they have the same *sensory-perceptual* phenomenology. Skeptics about the existence of proprietary language-understanding phenomenology typically respond by claiming that although the Chinese understander probably has different phenomenology than the non-understander, the differences can all be explained as a matter of different non-cognitive phenomenology: the spoken words very likely generate in the Chinese-understander certain content-appropriate mental images, and/or certain content-appropriate emotional responses, that will not arise in the person who hears the spoken Chinese words but does not understand them.[7]

This move is blocked, in the case of the phenomenal contrast argument employing scenario 5. Items (2) and (3) in the scenario guarantee, by stipulation, that this guy (the guy in the scenario) has exactly the same non-cognitive phenomenology that is present in the other guy—no less and no more.

How else might the skeptic about cognitive phenomenology try to explain the difference between this guy and the other guy? The move one would expect is an appeal to Ned Block's influential distinction between access consciousness and phenomenal consciousness as follows:

> The exercise of language understanding consists in undergoing certain kinds of cognitive states that are access conscious but lack any proprietary phenomenal character. The key difference between this guy and the other guy is that this guy fails to undergo any such access-conscious states, whereas the other guy undergoes lots of them. (Likewise, mutatis mutandis, for the differences in memory experience between this guy and that guy: these are all differences in what's access conscious, not differences in phenomenology.)[8]

But there are two reasons, I submit, why one should find this move unsatisfactory and unpersuasive. First, it seems intuitively very clear that the respective mental differences between this guy and the other guy concern the *intrinsic character* of certain mental states of this guy and the other guy, respectively—differences in how these states are *directly experienced*. Yet, any state that is merely access conscious, but not phenomenally conscious, has a mental essence that is completely functional and relational: its being access-conscious is entirely a matter of the effects it produces and is disposed to produce, by itself or in combination with other mental states. It therefore cannot manifest itself *directly* in experience at all—unlike phenomenally conscious states, which have intrinsic phenomenal character that is experientially self-presenting. Rather, it can only manifest itself indirectly, via *the phenomenology* that it (perhaps in combination with other mental states) causally generates—including sensory-perceptual and kinesthetic phenomenology whose content involves what one's own body is doing.[9] If there is no such thing as cognitive phenomenology, therefore, then the intrinsic character of the experiences of the guy in scenario 5 would turn out to be no different than the intrinsic character of the other guy's experiences! After all, differences in the intrinsic character of experience are *phenomenal* differences, and *ex hypothesi* the two guys' mental lives are phenomenally exactly the same with respect to all the kinds of phenomenal character that the cognitive-phenomenology denier acknowledges. So, even though the cognitive-phenomenology skeptic is appealing to the contention that this guy's conscious mental life differs from the other guy's mental life with respect to access-conscious mental states, nevertheless the skeptic must still embrace the grossly counterintuitive claim that this guy's mental life is *intrinsically* experientially exactly like the other guy's mental life.

The second reason to repudiate the "mere access consciousness" reply to my phenomenal-contrast argument is that the cognitive-phenomenology skeptic actually has no legitimate basis for claiming that this guy and the other guy differ with respect to their access-conscious mental states. For, in scenario 5 the MPS device is causally functionally interconnected with this guy's brain in such a way that the total system comprising this guy and the device undergoes internal states, sensory-input states, and behavioral states that collectively exhibit a causal-functional profile that exactly duplicates the causal-functional profile exhibited by the other guy's conscious internal states, sensory-input states, and behavioral states. But if indeed this guy and the other guy are mentally just alike phenomenally (as the skeptic about cognitive phenomenology is committed to saying about scenario 5), then such exact duplication of causal-functional mental profile means that this guy and the other guy are exactly alike not only with respect to their phenomenally conscious mental states but also with respect to the *full range* of their conscious mental states, both phenomenally conscious and access conscious! This guy's total conscious mental life exactly matches the total conscious mental life of the other guy because (i) this guy and the other guy supposedly are exactly alike with respect to their phenomenology, and (ii) the MPS device is integrated with this guy's brain-cum-body in such a way that the causal-functional profile of states that occur in this guy's brain-cum-body-cum-MPS-device constitutes an *alternative implementation* of the very same causal-functional mental profile that is implemented, in the other guy, entirely within the other guy's brain-cum-body.[10] It would be objectionable "implementation chauvinism" for the skeptic about cognitive phenomenology to deny this, and to embrace instead the claim that the goings-on in the MPS device are not part of this guy's conscious mental life.

The upshot is this. The only plausible explanation of the differences between the respective conscious mental lives of this guy and the other guy is that this guy lacks Chinese language-understanding phenomenology (and memory phenomenology), whereas the other guy possesses them both.

Conclusion

Phenomenal consciousness, comprising those kinds of mental state that have a distinctive, proprietary, and individuative "what it is like" aspect to them, is philosophically mysterious. It gives rise to what Joseph Levine calls the "explanatory gap" and David Chalmers calls the "hard problem," consisting of questions like the following.[11] Why should it be that such-and-such physical state, or functional state, or functional-state-cum-physical-realization, has *this* experientially distinctive phenomenal character (e.g., visually presenting an apple as *looking green*), rather than having *that* phenomenal character (say, visually presenting the apple as *looking red*), or rather than having no phenomenal character at all? Can phenomenal consciousness be smoothly integrated into a broadly naturalistic—perhaps even materialist—metaphysics, and if so, how?

Intentionality, in thought and in public language, often has been thought to be largely separate from phenomenal consciousness, and also philosophically less mysterious—even among philosophers who accept the claim that phenomenal consciousness poses a "hard problem." This is because functionalist orthodoxy about intentional mental states has remained dominant, along with a widespread tendency to think that phenomenal consciousness only constitutes a relatively circumscribed portion of one's conscious-as-opposed-to-unconscious mental life. Prototypically intentional mental states—e.g., occurrent propositional attitudes—have been widely thought to lack any phenomenal character that is distinctive, proprietary, and individuative. Also, it has been widely thought that such states possess full fledged intentionality, and do so solely by virtue of their *functional roles*—roles that perhaps incorporate various constitutive connections (e.g., causal, and/or covariational, and/or evolutionary-historical) to the cognitive agent's wider environment. (Functionalist orthodoxy about mental intentionality has gone strongly externalist.)

If one embraces this recently dominant conception of intentional mental states, then one is apt to think that suitably sophisticated robots could undergo such states, solely by virtue of having the right kind of functional architecture. Even a robot that had no phenomenology at all—a "zombie robot"—could be a full-fledged *cognitive* agent, with propositional-attitude mental states that possess full-fledged, nonderivative intentionality. (It is worth recalling that Hilary Putnam originated functionalism in philosophy of mind, and that his earliest writings on the topic were couched in terms of Turing machines and probabilistic automata, and were entitled "Minds and Machines" and "The Mental Life of Some Machines."[12]

In my view, there is indeed a hard problem of phenomenal consciousness, and an accompanying explanatory gap. But the functionalist conception of mental intentionality, still dominant in philosophy of mind, is profoundly mistaken. Searle was right about the guy in the Chinese Room, and about the whole guy-in-room system, and about the guy-guided robot. The reason he was right—the reason why neither the guy, nor the guy/room system, nor the guy-guided robot, understands Chinese—is that they all lack the distinctive, proprietary, individuative *phenomenology* that constitutes genuine Chinese language-understanding. And this point is highly generalizable: full-fledged mental intentionality is phenomenal intentionality. This means, among other things, that zombie robots would have no genuine mental life at all. Perhaps robots that really think are possible, but if so it would not be solely because of their functional architecture. Rather, in order to be real thinkers, they would have to undergo cognitive phenomenology—as do we humans.

I recognize that this approach to mental intentionality makes the hard problem more pervasive than it is often thought to be, and even harder. And, for what it's worth, I continue to be a "wannabe materialist" about the mind and its place in nature—although I have little idea what an adequate version of materialism would look like. But one should not mischaracterize mental intentionality because one would like to naturalize it.[13]

Notes

1. See, for instance, Horgan and Tienson, "Intentionality of Phenomenology" and "Phenomenology of Embodied Agency"; Horgan, Tienson, and Graham, "Phenomenology of First-Person Agency," "Phenomenal Intentionality," and "Internal-World Skepticism"; Graham, Horgan, and Tienson, "Consciousness and Intentionality" and "Phenomenology, Intentionality, and the Unity of Mind"; Horgan, "From Agentive Phenomenology to Cognitive Phenomenology" and "Phenomenology of Agency and Freedom"; and Horgan and Graham, "Phenomenal Intentionality and Content Determinacy." Closely related writings by others include McGinn, *Mental Content*; Strawson, *Mental Reality*; Loar, "Phenomenal Intentionality as the Basis for Mental Content; Pitt, "Phenomenology of Cognition"; Siewert, *Significance of Consciousness*; Kriegel, *Sources of Intentionality*; and Kriegel, *Phenomenal Intentionality*.
2. This paper originated as a talk by the same title, presented at a symposium on machine consciousness at the 2012 meeting of the American Philosophical Association in Chicago, organized by David Anderson. Sections 3 and 4 largely coincide (with some additions) with material in Horgan, "Original Intentionality is Phenomenal Intentionality."
3. In Horgan and Tienson, *Connectionism*, John Tienson and I argue that human cognition is too subtle and too holistically information-sensitive to conform to programmable rules that operate on content-encoding structural features of mental representations. We describe a non-classical framework for cognitive science—inspired by connectionist modeling and the mathematics of dynamical systems theory—that we call "non-computational dynamical cognition."
4. My use of the locution "in virtue of," here and in the preceding paragraph, is meant to pick out a *conceptually constitutive* requirement for genuine language understanding. Features that together constitute a supervenience base for understanding-phenomenology—where the strength of the modal connection between the subvenient features and the supervenient phenomenal features is either nomic necessity or metaphysical necessity—thus do not bear the intended kind of *in-virtue-of* relation to genuine language understanding.
5. On the method of phenomenal contrast, see Siegel, "Which Properties are Represented in Perception?" and *Contents of Visual Experience*. I am using the term "experience" in a way that deliberately brackets the issue of how extensive phenomenal character is. Experience comprises those aspects of mentality that are conscious-as-opposed-to-unconscious; this leaves open how much of what is in experience is *phenomenally* conscious, as opposed to merely being "access conscious" (cf. Block, "Function of Consciousness"). On my usage, the agreed-upon *experiential* difference that feeds into a phenomenal contrast argument need not be one that both parties would happily call a *phenomenal* difference. Rather, the claim will be that a posited phenomenal difference best explains the agreed-upon experiential difference.
6. I give a somewhat similar argument, focused around aspects of agentive phenomenology, in Horgan, "From Agentive Phenomenology to Cognitive Phenomenology."

7. This response assumes that content-appropriate emotional responses would have phenomenal character but not *cognitive* phenomenal character. That assumption seems dubious; it suggests, for instance, that the phenomenal character of the experience of getting a specific joke is a generic, non-intentional, mirthfulness phenomenology—rather than being the what-it's-like of *content-specific* mirthfulness. But I am granting, for the sake of argument, the (dubious) assumption that the phenomenal character of emotions that would be apt responses to language one understands would be *non-cognitive* phenomenal character, divorceable from the content of what is understood.
8. Block, "Function of Consciousness."
9. Once this fact is fully appreciated, it becomes very plausible that states that are only access conscious in Block's sense, without possessing proprietary phenomenal character, are not really conscious *in the pre-theoretic sense* at all. But my argument does not require this to be so.
10. What about those states, subserved within this guy's brain, of the kind I described as being experiences as-of hearing meaningless-seeming noises, and experiences as-of having spontaneous desires to spontaneously move one's body in various pointless-seeming ways? Well, I myself claim that these experiences have cognitive-phenomenal character—and, indeed, very different cognitive-phenomenal character than is present in the other guy's mental life. But the skeptic about cognitive phenomenology must deny that these brain-subserved states have any inherent phenomenal character, and also must regard them as mere accompaniments to this guy's concurrent non-cognitive phenomenal states. So, as far as I can see, the cognitive-phenomenology skeptic has no principled basis for treating these states as *mental* at all; rather, evidently they should be treated as mere sub-mental *causal intermediaries* between (i) states of the MPS device that implement certain merely-access-conscious states in this guy's causal-functional mental profile, and (ii) states of this guy's brain-cum-body that involve the other aspects of this guy's conscious mental life—viz., sensory states and other non-cognitive phenomenal states, brain-subserved merely-access-conscious states, and behaviors.
11. Levine, "Materialism and Qualia"; Levine, *Purple Haze*; Chalmers, *Conscious Mind*.
12. Putnam, "Minds and Machines"; Putnam, "Mental Life of Some Machines."
13. My thanks to the audience at the symposium on machine consciousness at the 2012 Central Division APA Meeting, and to Steven Gubka, Rachel Schneebaum, and John Tienson for helpful comments and discussion. My thanks to Peter Boltuc, a participant in the symposium, for inviting me to contribute this paper to the *Newsletter on Philosophy and Computers*.

Bibliography

Block, N. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (1995): 227–87.

Chalmers, D. *The Conscious Mind*. Oxford: OUP, 1996.

Graham, G., T. Horgan, and J. Tienson. *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press, 1996.

———. "Consciousness and Intentionality." In *The Blackwell Companion to Consciousness*, edited by M. Valmans and S. Schneider. Oxford: Blackwell Publishing, 2007.

———. "Phenomenology, Intentionality, and the Unity of Mind." In *The Oxford Handbook of Philosophy of Mind*, edited by B. McLaughlin, A. Beckermann, and S. Walter. Oxford: Oxford University Press, 2009.

Kriegel, U. *The Sources of Intentionality*. Oxford: Oxford University Press, 2011.

———. *Phenomenal Intentionality: New Essays*. Oxford: Oxford University Press, 2012.

Horgan, T. "From Agentive Phenomenology to Cognitive Phenomenology: A Guide for the Perplexed." In *Cognitive Phenomenology*, edited by T. Bayne and M. Montague. Oxford, UK: Oxford University Press, 2011.

———. "The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits." *Humana Mente* 15 (2011): 77–97.

———. "Original Intentionality is Phenomenal Intentionality." *The Monist* 96 (2013): 232–51.

Horgan, T. and G. Graham. "Phenomenal Intentionality and Content Determinacy." In *Prospects for Meaning*, edited by R. Schantz. Berlin: de Gruyter, 2012.

Horgan, T. and J. Tienson. "The Intentionality of Phenomenology and the Phenomenology of Intentionality." In *Philosophy of Mind: Classical and Contemporary Readings*, edited by D. Chalmers. Oxford, UK: Oxford University Press, 2002.

———. "The Phenomenology of Embodied Agency." In *A Explicacao da Interpretacao Humana: The Explanation of Human Interpretation. Proceedings of the Conference Mind and Action III—May 2001*, edited by M. Saagua and F. de Ferro. Lisbon: Edicoes Colibri, 2005.

———. *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press, 1996.

Horgan, T., J. Tienson, and G. Graham. "The Phenomenology of First-Person Agency." In *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, edited by S. Walter and H. D. Heckmann. Exeter: Imprint Academic, 2003.

———. "Phenomenal Intentionality and the Brain in a Vat." In *The Externalist Challenge*, edited by R. Schantz. Berlin: de Gruyter, 2004.

———. "Internal-World Skepticism and the Self-Presentational Nature of Phenomenal Consciousness." In *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium*, edited by M. Reicher and J. Marek. Vienna: Obv & hpt, 2005. Also in U. Kriegel and K. Williford, eds., *Self-Representational Approaches to Consciousnes*. Cambridge, MA: MIT Press, 2006.

Levine, J. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (1988): 354–61.

———. *Purple Haze*. Oxford: Oxford University Press, 2001.

Loar, B. "Phenomenal Intentionality as the Basis for Mental Content." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by M. Hahn and B. Ramberg. Cambridge, MA: MIT Press, 1987.

McGinn, C. *Mental Content*. Oxford, UK: Blackwell Publishing, 1989.

Pitt, D. "The Phenomenology of Cognition: Or What is It Like to Think that P?," *Philosophy and Phenomenological Research* 69 (2004): 1–36.

Putnam, H. "Minds and Machines." In *Dimensions of Mind*, edited by S. Hook. New York: New York University Press, 1960.

———. "The Mental Life of Some Machines." In *Intentionality, Minds, and Perception*, edited by H. Casteneda. Detroit: Wayne State University Press, 1967.

Searle, J. "Minds, Brains, and Programs." *The Behavioral and Grain Sciences* 3 (1980): 417–57.

Siegel, S. "Which Properties are Represented in Perception?" In *Perceptual Experience*, edited by T. Gendler Szabo and J. Hawthorne. Oxford, UK: Oxford University Press, 2006.

———. *The Contents of Visual Experience*. Oxford, UK: Oxford University Press, 2010.

Siewert, C. *The Significance of Consciousness*. Princeton: Princeton University Press, 1998.

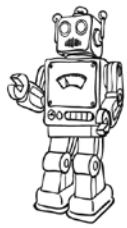
Strawson, G. *Mental Reality*. Cambridge, MA: MIT Press, 1994.

WILL A MACHINE EVER BE CONSCIOUS?

name: rock
genre: inorganic
living: no
intelligent: no
conscious: no

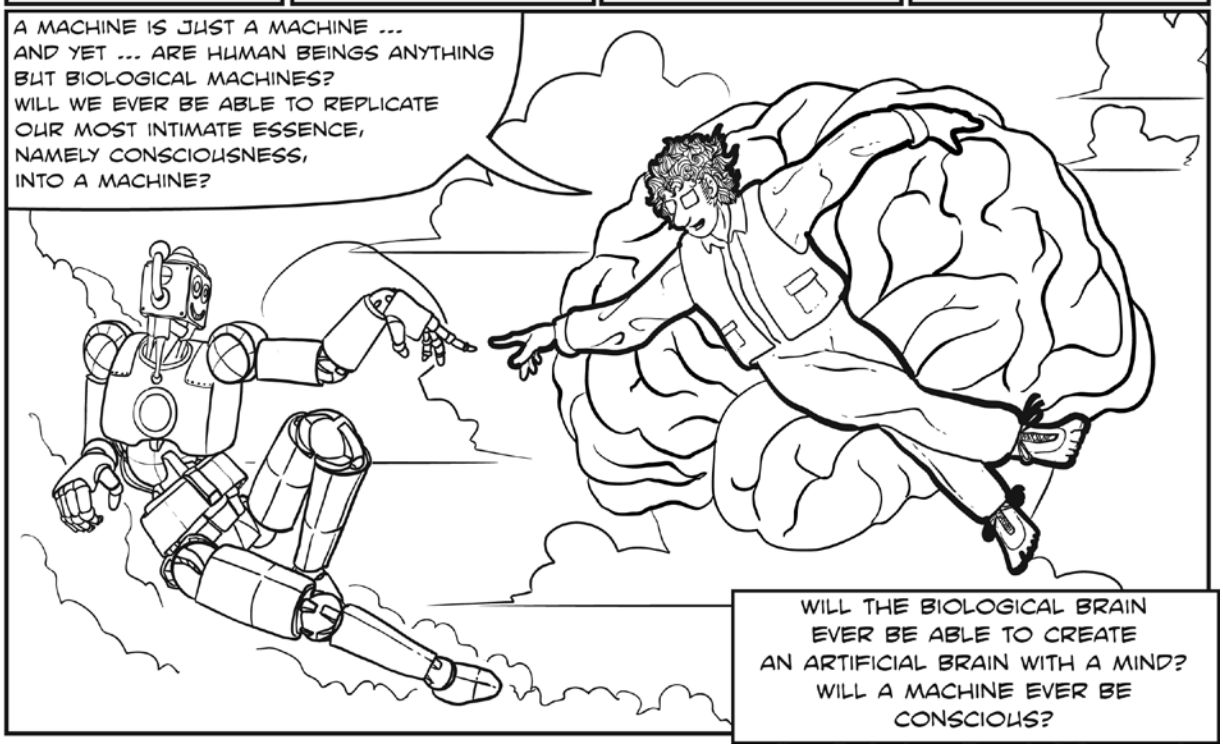
name: Lucky
genre: animal
living: yes
intelligent: enough
conscious: yes

name: Rick
genre: human
living: yes
intelligent: ?
conscious: yes

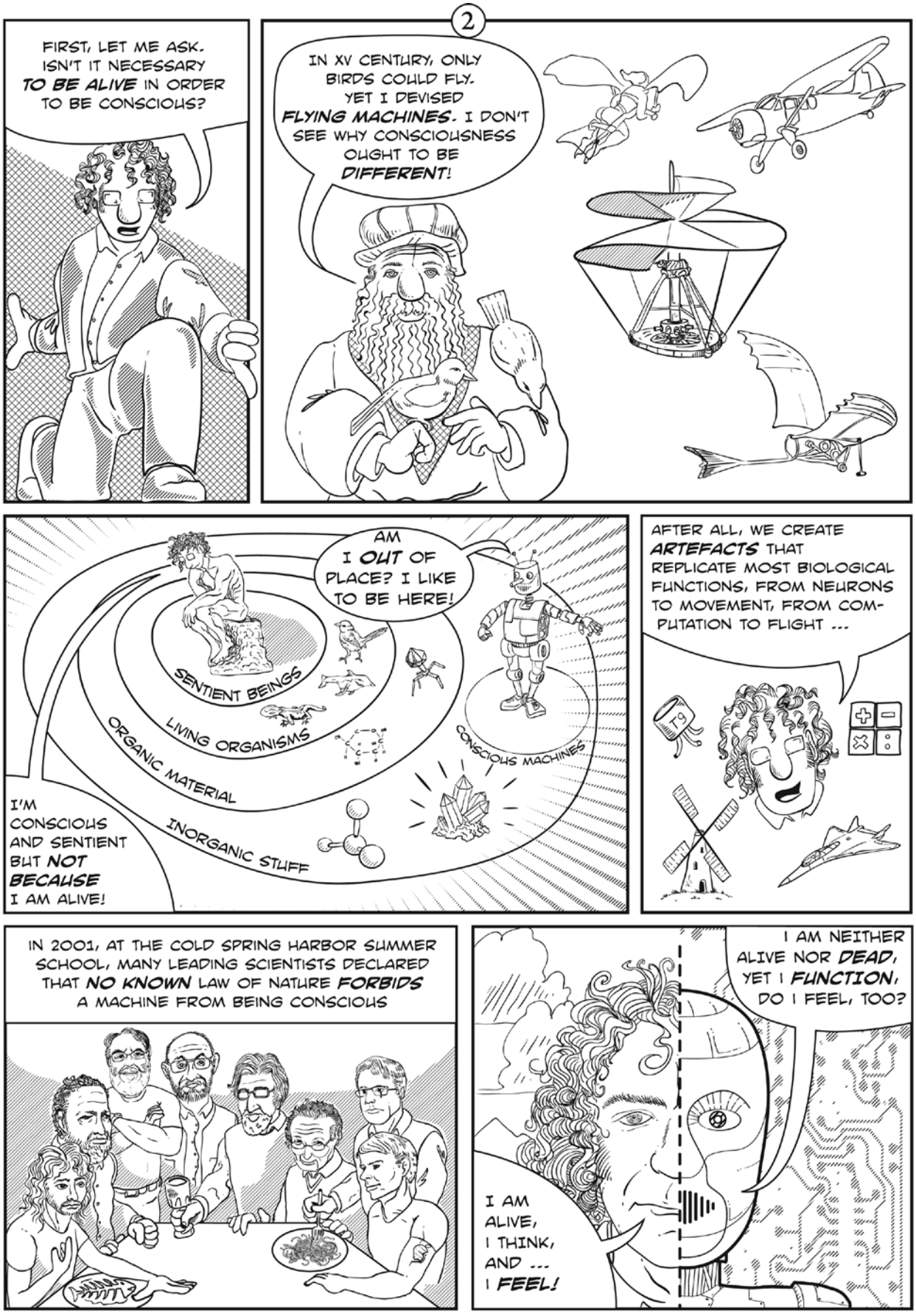


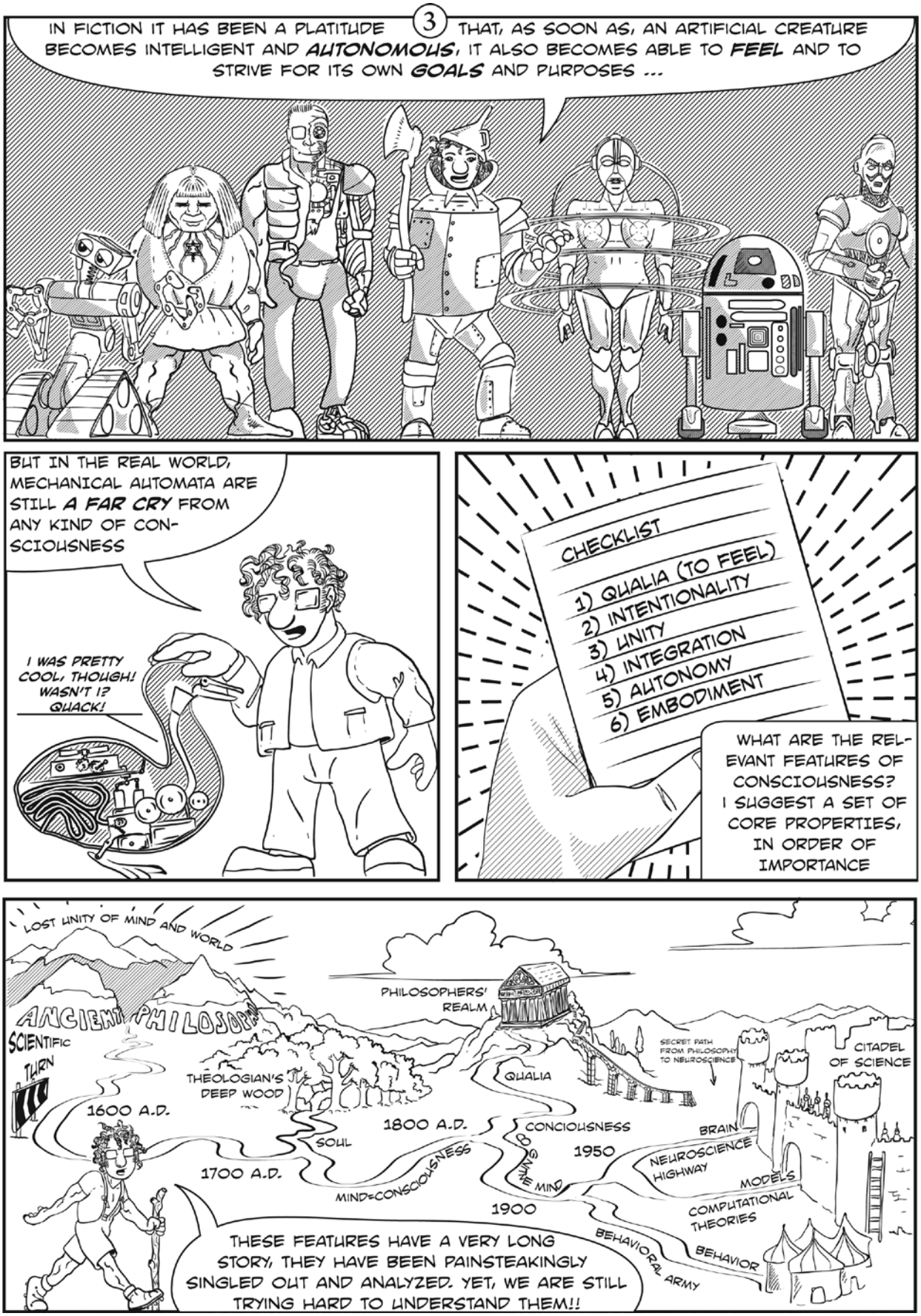
name: Buck
genre: artificial
living: no
intelligent: a little
conscious: not yet

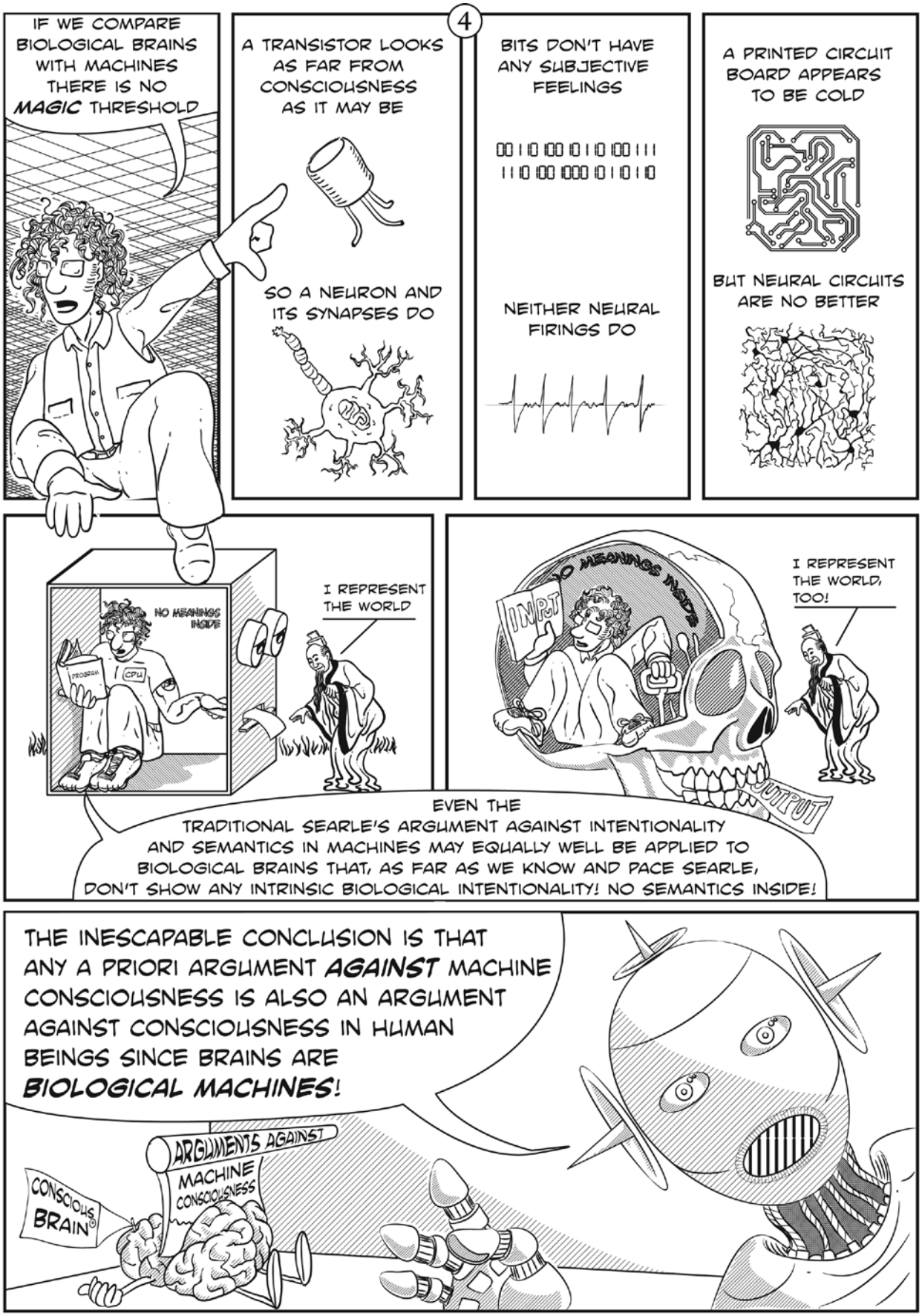
A MACHINE IS JUST A MACHINE ... AND YET ... ARE HUMAN BEINGS ANYTHING BUT BIOLOGICAL MACHINES? WILL WE EVER BE ABLE TO REPLICATE OUR MOST INTIMATE ESSENCE, NAMELY CONSCIOUSNESS, INTO A MACHINE?

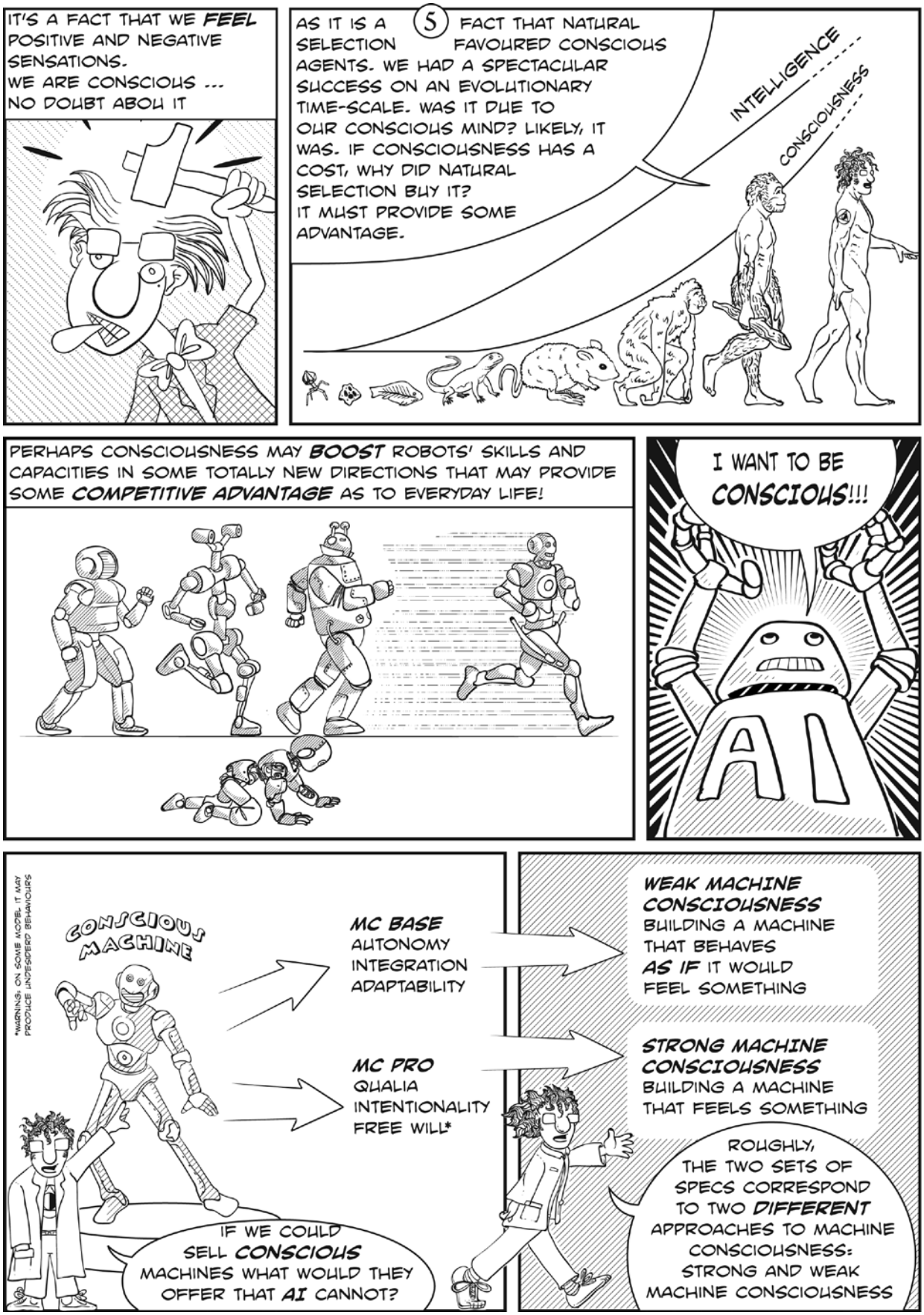


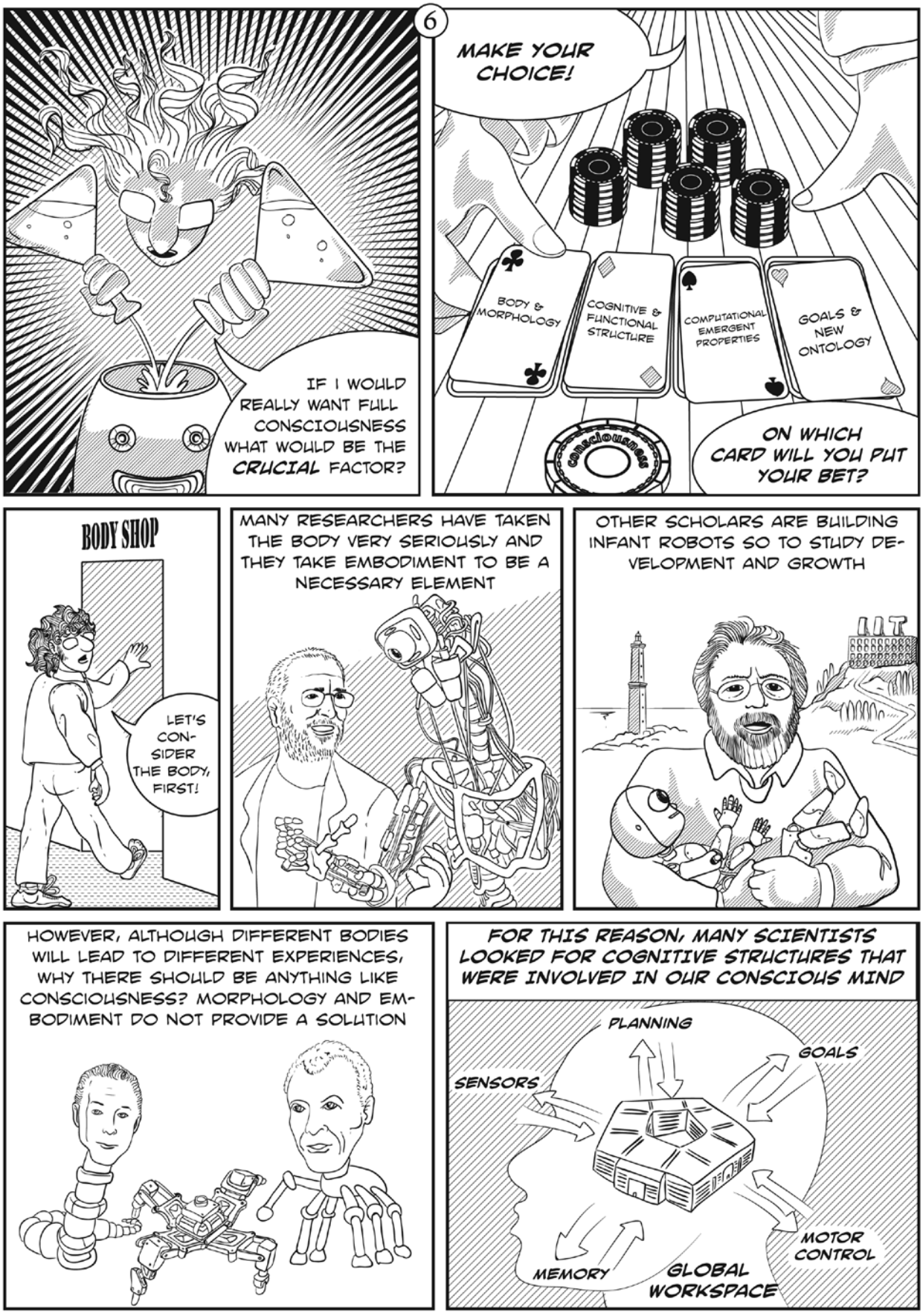
WILL THE BIOLOGICAL BRAIN EVER BE ABLE TO CREATE AN ARTIFICIAL BRAIN WITH A MIND? WILL A MACHINE EVER BE CONSCIOUS?

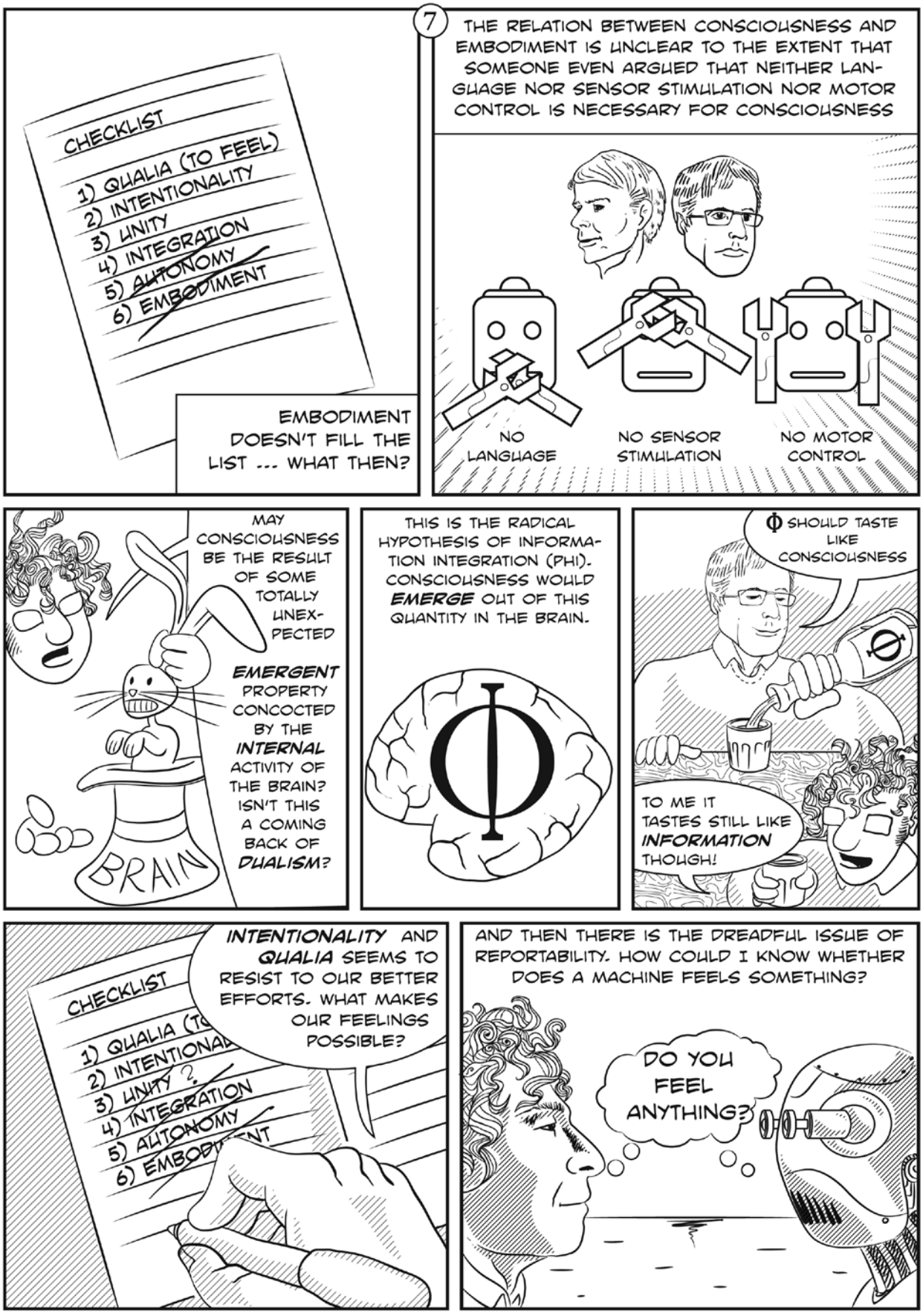

FIRST, LET ME ASK. ISN'T IT NECESSARY TO BE ALIVE IN ORDER TO BE CONSCIOUS?
2
IN XV CENTURY, ONLY BIRDS COULD FLY. YET I DEVISED FLYING MACHINES. I DON'T SEE WHY CONSCIOUSNESS OUGHT TO BE DIFFERENT!
AM I OUT OF PLACE? I LIKE TO BE HERE!
SENTIENT BEINGS
LIVING ORGANISMS
ORGANIC MATERIAL
INORGANIC STUFF
CONSCIOUS MACHINES
I'M CONSCIOUS AND SENTIENT BUT NOT BECAUSE I AM ALIVE!
AFTER ALL, WE CREATE ARTEFACTS THAT REPLICATE MOST BIOLOGICAL FUNCTIONS, FROM NEURONS TO MOVEMENT, FROM COMPUTATION TO FLIGHT ...
IN 2001, AT THE COLD SPRING HARBOR SUMMER SCHOOL, MANY LEADING SCIENTISTS DECLARED THAT NO KNOWN LAW OF NATURE FORBIDS A MACHINE FROM BEING CONSCIOUS
I AM NEITHER ALIVE NOR DEAD, YET I FUNCTION. DO I FEEL, TOO?
I AM ALIVE, I THINK, AND ... I FEEL!

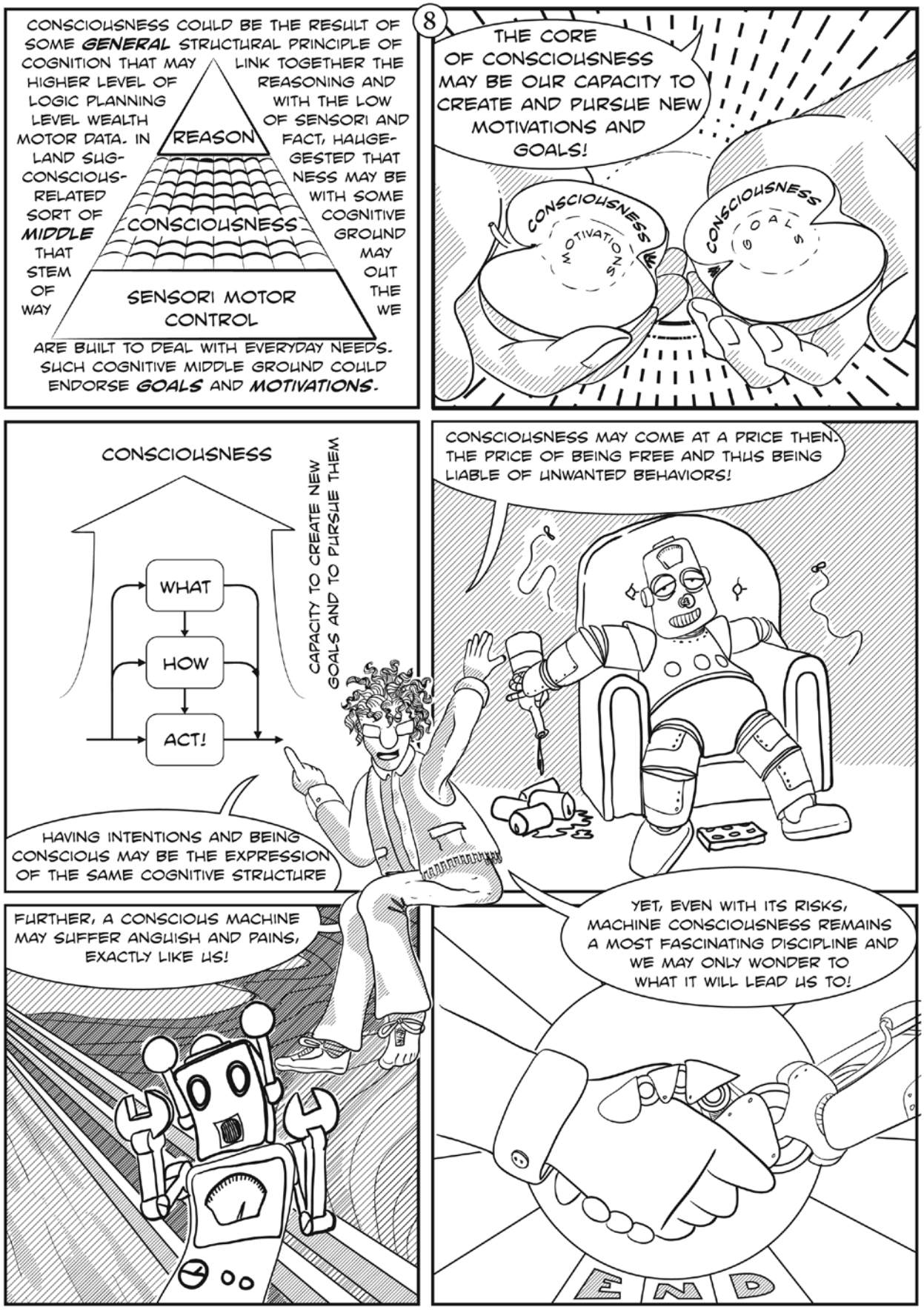

3
IN FICTION IT HAS BEEN A PLATITUDE THAT, AS SOON AS, AN ARTIFICIAL CREATURE BECOMES INTELLIGENT AND AUTONOMOUS, IT ALSO BECOMES ABLE TO FEEL AND TO STRIVE FOR ITS OWN GOALS AND PURPOSES ...
BUT IN THE REAL WORLD, MECHANICAL AUTOMATA ARE STILL A FAR CRY FROM ANY KIND OF CONSCIOUSNESS
I WAS PRETTY COOL, THOUGH! WASN'T I? QUACK!
CHECKLIST
1) QUALIA (TO FEEL)
2) INTENTIONALITY
3) UNITY
4) INTEGRATION
5) AUTONOMY
6) EMBODIMENT
WHAT ARE THE RELEVANT FEATURES OF CONSCIOUSNESS? I SUGGEST A SET OF CORE PROPERTIES, IN ORDER OF IMPORTANCE
LOST UNITY OF MIND AND WORLD
ANCIENT PHILOSOPHY
SCIENTIFIC TURN
1600 A.D.
1700 A.D.
1800 A.D.
1900
1950
THEOLOGIAN'S DEEP WOOD
PHILOSOPHERS' REALM
SOUL
MIND=CONSCIOUSNESS
QUALIA
CONCIOUSNESS
ONE MIND
SECRET PATH FROM PHILOSOPHY TO NEUROSCIENCE
CITADEL OF SCIENCE
BRAIN
NEUROSCIENCE HIGHWAY
MODELS
COMPUTATIONAL THEORIES
BEHAVIOR
BEHAVIORAL ARMY
THESE FEATURES HAVE A VERY LONG STORY, THEY HAVE BEEN PAINSTEAKINGLY SINGLED OUT AND ANALYZED. YET, WE ARE STILL TRYING HARD TO UNDERSTAND THEM!!


IF WE COMPARE BIOLOGICAL BRAINS WITH MACHINES THERE IS NO MAGIC THRESHOLD
4
A TRANSISTOR LOOKS AS FAR FROM CONSCIOUSNESS AS IT MAY BE
SO A NEURON AND ITS SYNAPSES DO
BITS DON'T HAVE ANY SUBJECTIVE FEELINGS
NEITHER NEURAL FIRINGS DO
A PRINTED CIRCUIT BOARD APPEARS TO BE COLD
BUT NEURAL CIRCUITS ARE NO BETTER
NO MEANINGS INSIDE
PROGRAM
CPU
I REPRESENT THE WORLD
NO MEANINGS INSIDE
INPUT
OUTPUT
I REPRESENT THE WORLD, TOO!
EVEN THE TRADITIONAL SEARLE'S ARGUMENT AGAINST INTENTIONALITY AND SEMANTICS IN MACHINES MAY EQUALLY WELL BE APPLIED TO BIOLOGICAL BRAINS THAT, AS FAR AS WE KNOW AND PACE SEARLE, DON'T SHOW ANY INTRINSIC BIOLOGICAL INTENTIONALITY! NO SEMANTICS INSIDE!
THE INESCAPABLE CONCLUSION IS THAT ANY A PRIORI ARGUMENT AGAINST MACHINE CONSCIOUSNESS IS ALSO AN ARGUMENT AGAINST CONSCIOUSNESS IN HUMAN BEINGS SINCE BRAINS ARE BIOLOGICAL MACHINES!
ARGUMENTS AGAINST MACHINE CONSCIOUSNESS
CONSCIOUS BRAIN©


IT'S A FACT THAT WE FEEL POSITIVE AND NEGATIVE SENSATIONS.
WE ARE CONSCIOUS ... NO DOUBT ABOU IT
AS IT IS A 5 FACT THAT NATURAL SELECTION FAVOURED CONSCIOUS AGENTS. WE HAD A SPECTACULAR SUCCESS ON AN EVOLUTIONARY TIME-SCALE. WAS IT DUE TO OUR CONSCIOUS MIND? LIKELY, IT WAS. IF CONSCIOUSNESS HAS A COST, WHY DID NATURAL SELECTION BUY IT? IT MUST PROVIDE SOME ADVANTAGE.
INTELLIGENCE
CONSCIOUSNESS
PERHAPS CONSCIOUSNESS MAY BOOST ROBOTS' SKILLS AND CAPACITIES IN SOME TOTALLY NEW DIRECTIONS THAT MAY PROVIDE SOME COMPETITIVE ADVANTAGE AS TO EVERYDAY LIFE!
I WANT TO BE CONSCIOUS!!!
AI
CONSCIOUS MACHINE
*WARNING: ON SOME MODEL IT MAY PROVE UNDESIRED BEHAVIOURS
MC BASE
AUTONOMY
INTEGRATION
ADAPTABILITY
MC PRO
QUALIA
INTENTIONALITY
FREE WILL*
IF WE COULD SELL CONSCIOUS MACHINES WHAT WOULD THEY OFFER THAT AI CANNOT?
WEAK MACHINE CONSCIOUSNESS
BUILDING A MACHINE THAT BEHAVES AS IF IT WOULD FEEL SOMETHING
STRONG MACHINE CONSCIOUSNESS
BUILDING A MACHINE THAT FEELS SOMETHING
ROUGHLY, THE TWO SETS OF SPECS CORRESPOND TO TWO DIFFERENT APPROACHES TO MACHINE CONSCIOUSNESS: STRONG AND WEAK MACHINE CONSCIOUSNESS


6
IF I WOULD REALLY WANT FULL CONSCIOUSNESS WHAT WOULD BE THE CRUCIAL FACTOR?
MAKE YOUR CHOICE!
BODY & MORPHOLOGY
COGNITIVE & FUNCTIONAL STRUCTURE
COMPUTATIONAL EMERGENT PROPERTIES
GOALS & NEW ONTOLOGY
consciousness
ON WHICH CARD WILL YOU PUT YOUR BET?
BODY SHOP
LET'S CONSIDER THE BODY, FIRST!
MANY RESEARCHERS HAVE TAKEN THE BODY VERY SERIOUSLY AND THEY TAKE EMBODIMENT TO BE A NECESSARY ELEMENT
OTHER SCHOLARS ARE BUILDING INFANT ROBOTS SO TO STUDY DEVELOPMENT AND GROWTH
MIT
HOWEVER, ALTHOUGH DIFFERENT BODIES WILL LEAD TO DIFFERENT EXPERIENCES, WHY THERE SHOULD BE ANYTHING LIKE CONSCIOUSNESS? MORPHOLOGY AND EMBODIMENT DO NOT PROVIDE A SOLUTION
FOR THIS REASON, MANY SCIENTISTS LOOKED FOR COGNITIVE STRUCTURES THAT WERE INVOLVED IN OUR CONSCIOUS MIND
PLANNING
GOALS
SENSORS
MOTOR CONTROL
MEMORY
GLOBAL WORKSPACE


CHECKLIST
1) QUALIA (TO FEEL)
2) INTENTIONALITY
3) UNITY
4) INTEGRATION
5) AUTONOMY
6) EMBODIMENT
EMBODIMENT DOESN'T FILL THE LIST --- WHAT THEN?
7
THE RELATION BETWEEN CONSCIOUSNESS AND EMBODIMENT IS UNCLEAR TO THE EXTENT THAT SOMEONE EVEN ARGUED THAT NEITHER LANGUAGE NOR SENSOR STIMULATION NOR MOTOR CONTROL IS NECESSARY FOR CONSCIOUSNESS
NO LANGUAGE
NO SENSOR STIMULATION
NO MOTOR CONTROL
MAY CONSCIOUSNESS BE THE RESULT OF SOME TOTALLY UNEXPECTED EMERGENT PROPERTY CONCOCTED BY THE INTERNAL ACTIVITY OF THE BRAIN? ISN'T THIS A COMING BACK OF DUALISM?
BRAIN
THIS IS THE RADICAL HYPOTHESIS OF INFORMATION INTEGRATION (PHI). CONSCIOUSNESS WOULD EMERGE OUT OF THIS QUANTITY IN THE BRAIN.
Φ
Φ SHOULD TASTE LIKE CONSCIOUSNESS
TO ME IT TASTES STILL LIKE INFORMATION THOUGH!
CHECKLIST
INTENTIONALITY AND QUALIA SEEMS TO RESIST TO OUR BETTER EFFORTS. WHAT MAKES OUR FEELINGS POSSIBLE?
AND THEN THERE IS THE DREADFUL ISSUE OF REPORTABILITY. HOW COULD I KNOW WHETHER DOES A MACHINE FEELS SOMETHING?
DO YOU FEEL ANYTHING?


CONSCIOUSNESS COULD BE THE RESULT OF SOME GENERAL STRUCTURAL PRINCIPLE OF COGNITION THAT MAY LINK TOGETHER THE HIGHER LEVEL OF REASONING AND LOGIC PLANNING WITH THE LOW LEVEL WEALTH OF SENSORI AND MOTOR DATA. IN FACT, HAUGE- LAND SUG- GESTED THAT CONSCIOUS- NESS MAY BE RELATED WITH SOME SORT OF COGNITIVE MIDDLE GROUND THAT MAY STEM OUT OF THE WAY WE
REASON
CONSCIOUSNESS
SENSORI MOTOR CONTROL
ARE BUILT TO DEAL WITH EVERYDAY NEEDS. SUCH COGNITIVE MIDDLE GROUND COULD ENDORSE GOALS AND MOTIVATIONS.
8
THE CORE OF CONSCIOUSNESS MAY BE OUR CAPACITY TO CREATE AND PURSUE NEW MOTIVATIONS AND GOALS!
CONSCIOUSNESS
MOTIVATIONS
CONSCIOUSNESS
GOALS
CONSCIOUSNESS
WHAT
HOW
ACT!
CAPACITY TO CREATE NEW GOALS AND TO PURSUE THEM
CONSCIOUSNESS MAY COME AT A PRICE THEN. THE PRICE OF BEING FREE AND THUS BEING LIABLE OF UNWANTED BEHAVIORS!
HAVING INTENTIONS AND BEING CONSCIOUS MAY BE THE EXPRESSION OF THE SAME COGNITIVE STRUCTURE
FURTHER, A CONSCIOUS MACHINE MAY SUFFER ANGUISH AND PAINS, EXACTLY LIKE US!
YET, EVEN WITH ITS RISKS, MACHINE CONSCIOUSNESS REMAINS A MOST FASCINATING DISCIPLINE AND WE MAY ONLY WONDER TO WHAT IT WILL LEAD US TO!
END

My Avatar, My Choice! How Might We Make a Strong Case for the Special Moral Status of Avatars?

Roxanne Marie Kurtz
University of Illinois, Springfield

In 1993, Julian Dibbell described a case of virtual rape. A text-based computer avatar, Mr. Bungle, sexually brutalized his avatar victims. He began with "more or less conventional" sex acts and went on to cause one user "to violate herself with a piece of kitchen cutlery." Dibble continued to describe the trauma of the victims, the drivers of the avatars in the non-virtual world, citing their "posttraumatic tears."[1]

Virtual rape is an attack in which one avatar engages in what is understood to be nonconsensual sexual behavior with another avatar.[2] Clearly, the harm of virtual rape is not morally equivalent to the harm of physical rape. The harms differ in both degree and kind. Nevertheless, something deeper is going on than disturbing pretend play with computer-realized dolls.[3] I think we should accept the authenticity of the traumatic experiences of victims of virtual rape at face value (or at least I invite the reader to do so for this paper). In such cases, morally problematic violations of the driver—the person running the avatar—can occur. To be clear: my concern here is with *harm to the person* who runs an avatar, not harm to an avatar. How might we explain moral concerns involving such harms?

I think one sort of explanation is particularly unappealing. We might, upon recognizing the legitimacy of property rights in the virtual world, suggest that the moral problem of virtual rape reduces to the moral problem of property damage.[4] Maybe all that matters morally about avatars is that they are the drivers' personal property. But, if so, it seems we must reject the authenticity of the traumatic experiences of victims of virtual rape. On such a view, it would appear that they have suffered disproportionately to something that is indeed of no greater moral significance than very ugly play with dolls. To reduce virtual rape to mere property damage strips it of its moral significance in a way that I reject.

There are more promising paths to take in terms of finding a response that respects the moral harms that result from virtual rape. In a more or less straightforward manner, we might aim to account for the immorality of virtual rapes by classifying them as instances of hate speech or problematic speech acts or sexual harassment. If so, then we will encounter the kinds of moral debates that go with that moral territory. For instance, if we explain the moral harm of virtual rape as an instance of hate speech, we will face critics who dismiss the explanation on the grounds that purported victims may simply choose to ignore the speech with a click of the mute button.[5] Others who resist the "don't listen" defense of hate speech may similarly resist such a defense in the case of virtual rape. Or, some might question if the virtual behavior falls into the relevant category—one might freely assent that sexual harassment is problematic but deny that instances of sexual harassment can occur in the virtual world, a move others would counter. Arguments that aim to classify virtual rape as a case of hate speech, speech act, or sexual harassment may be the best way to explain the moral harm of virtual rape. But I do not seek to assess them here.

My focus is on what our arguments should look like if something more unusual is going on in cases of avatar violence. The wrongness of virtual rape suggests that avatars may have special moral significance in some way. I've wondered how we might take this idea seriously. Suppose we deny that the moral problem of virtual rape lies (solely) in concerns about property, speech, harassment, or other standard moral concerns. Suppose we think there are things it is worse to do to an avatar than a mere doll or puppet precisely because of the special moral status of the avatars. This is the sort of thinking that led me to ask the following question: How might we make the best case for the special moral status of computer-based avatars?

My thesis is that the best case for assigning special moral status to avatars will be grounded in an appeal to ways in which avatars are analogous to our bodies.

The challenge at hand

Suppose we support the claim below:

> **AvatarsMatter:** In general, the distinctive role of avatars in our lives confers upon them special moral status that distinguishes them from mere artifacts or mere property.

I do not seek to persuade you of the intuitive appeal of **AvatarsMatter**. Rather, my concern is that on the assumption we find it appealing, we ought to be able to explain why the moral status of avatars is generally different and greater than the moral status of puppets and dolls. Our task is to identify a promising justificatory strategy for **AvatarsMatter** that:

(1) Applies generally. We seek a way to motivate the idea that avatars have special moral status in everyday cases under ordinary conditions.[6]

(2) Involves the distinctive role of avatars in our lives. We do not want a story that works equally well for mere artifacts and mere property in standard cases.

(3) Demonstrates moral significance. We seek a reason to believe that not only is the moral status of avatars distinctive, but that it is greater than that of mere artifacts and mere property in standard cases.[7]

My response below to this challenge comes in two parts. First, I undertake a negative project to show that justifications that appeal to the psychological significance of avatars do not do a great job with respect to satisfying the above constraints. I then move onto a positive project in which I contend that the most promising line of justification is to argue for avatar control rights as analogous to bodily control rights.

Why is this of interest? To me, it is an example of emerging moral questions on the scene due to technology.[8] We have brought non-living things into the world that seem to have greater moral status than the kind of artifacts or property with which we are more used to dealing because of the ways they affect our basic interactions with the world. Thanks to technology, cochlear and corneal implants are not science fiction. Nor are bionic arms that we can move with our minds at a distance. Nor are eyeborg devices that allow a person to hear color as in the case of Neil Harbisson, a man who identifies as a cyborg who has monochromatic vision but now enjoys color through sound.[9] Nor are computer avatars that give us entry into the virtual world. And while fully functioning second bodies in the non-virtual world, such as Na'vi bodies that serve as organic avatars with human drivers in *Avatar* or the robot avatars also driven by humans in *Surrogates*,[10] still remain science fiction, neither seems too outlandish a possibility in the reasonably near future of humanity. We do, though, have virtual avatars already in the virtual world, driven by humans in the non-virtual world. Let us then begin to take notice of how technology might affect morality because of the important role such objects play in our lives.

Some time ago, we may have thought of virtual reality as no more than make-believe—a pretend world without significant moral relevance. In this paper, I take it for granted that we now understand that non-virtual and virtual reality are both parts of

reality proper.[11] I further assume that we take as given that right and wrong cross the virtual border, the concerns of morality reach into our virtual lives.[12]

We see this when we recognize that in real or virtual life, sometimes it is *morally permissible* to damage or harm things (living or otherwise). For instance, it is OK to put one's own doll in a trash compactor. It is OK to harm someone in self-defense. It is OK to hit someone in a boxing ring (perhaps). It is OK to delete one's own avatar. It is OK to "kill" an avatar in some virtual games.[13]

Likewise, we see that in real or virtual life, sometimes it is *morally impermissible* to damage or harm things (living or otherwise). It is not OK to put someone else's doll in a trash compactor without permission. It is not OK to harm someone for fun. It is not OK to hit someone to rob them. It is not OK to delete someone else's avatar without permission. It is not OK to "kill" an avatar in some contexts in virtual games.

At this point, my reader might ask, what is an avatar anyway? My response is that this is a tough question, and, I think, an interesting question. But, it is not a question of interest to me here. For our purposes, let us construe what it is to be an avatar rather vaguely as a virtual presence that allows us access to and causal power in virtual space. Perhaps an avatar is a collection of consecutive screen images, a bit of code, a fusion of the two. Or, it might be more broadly the sum of one's presence in the virtual world.[14] I invite the reader to understand "avatar" in whatever way she finds makes **AvatarsMatter** most plausible.[15] A driver is the person who runs the avatar.

The reader might also wonder just how important I think avatars are, morally speaking. Again, I refrain from taking a stand. I wish to allow a good deal of room for positions that vary on the significance they attribute to the moral status of avatars, but share the idea that there is something special or distinctive about them, morally speaking.

Part 1: The negative project

Some have argued that avatars have moral status because of their psychological importance, or simply because they fall under the category of personal property. I contend that neither approach to justifying special moral status for avatars meets the challenge described above. Let us consider these views in turn.

First, consider a position that ties avatar significance to psychological significance:[16]

> **Psychological thesis:** Avatars have special moral status because of the psychological attachment a driver has to her avatar.

I take it that "psychological" here has to do with our mental states involving beliefs and affect. Certainly a driver may form a significant psychological attachment to her avatar. Such attachments may vary across cases in ways that affect their moral import. Jessica Wolfendale recognizes this in her important paper, which influences my discussion below, though at times I am not quite sure which of the two interpretations of the psychological thesis I sketch below are her target.[17]

Consider that some people have psychological connections to avatars that create strong ties to self-identity in deeply important ways, perhaps as integral parts of their psychological self-image. Consider how this might work in analogous cases. In the movie *Avatar*, for instance, I would say that the drivers feel as though they just are their avatars or that they just count avatars as parts of their self-identity. It would be hard to imagine Jake experiencing love and intimacy with Neytiri in the *Avatar* world without the existence of profound connections between his physical avatar and psychological self. In a less fanciful case, consider the case of ventriloquists who seem to have a deep psychological bond to their puppets. The documentary *Dumbstruck* may be instructive here, in which some ventriloquists seem to deeply identify with their dummies, arguably to a degree that the dummies simply have become part of who they are, psychologically speaking. These suggestive quotes from reviewers support such a view:

> They've always known they'd be ventriloquists and have accepted that a life without a pint-sized, wooden partner wouldn't be worth living.[18]

> "Dumbstruck" keeps a handful of emotional tricks up its sleeve, even as it confirms the tough truths we may have suspected about ventriloquists—that they're lonely, unstable, shy and socially challenged people who need the help of an artificial friend if they're ever going to learn to stand up for themselves.[19]

> Watching these artists practicing in secret, listening to them use their dummies to say things they won't, it seems clear their art is the effect, not the cause. They learned to speak through others because nobody heard them.[20]

I think we really ought to allow that some people's psychological selves may include deep attachments to objects like physical avatars and dummies. Likewise, we really ought to acknowledge that some people's psychological selves may include deep attachments to objects like virtual avatars.

With such attachments, I grant that important moral concerns arise connected to those objects precisely because they are so important to the person's psychological well-being. I think we can agree that it is worse, morally speaking, to destroy the dummy of a ventriloquist who is profoundly attached to her dummy than it would be to destroy the dummy of a mere hobbyist who plays with it only each New Year's Eve right before the ball drops. The difference is not in the value of the dummy-in-itself, or in the value of personal property, but precisely in the moral requirement that we attend to personal contexts when considering how we may use objects to avoid needless harm to others. Because humans may form deep psychological attachment to objects, surely that gives us moral reasons to be careful with what we do with them. I take it that is the case regardless of our judgments as to the healthiness of the attachment. So, let us allow that likewise deep psychological attachment to a computer avatar gives us grounds for unusual moral concerns around an avatar in particular cases. But I think we ought to also recognize that such attachment is not the norm. We ought not attribute some special moral status for avatars that applies *generally* for such reasons. So on this way of understanding avatar attachment, we have not found resources to meet the generality constraint above.

Suppose instead that we understand the avatar attachment position as one that simply requires strong sentimental attachment to the avatar without the strong ties to psychological self. For instance, one might be sentimentally attached 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, or a wedding ring, or a favorite fossil. Surely loss or damage to an object to which we have strong sentimental attachments can result in emotional pain. As before, these attachments might make it the case that morality requires us to have more concern with our treatment of those objects than we have with similar objects to which no one has sentimental attachment (regardless of our judgments about the healthiness of the attachments). Likewise, surely the same would be true of computer avatars, notwithstanding their virtual nature. Moreover, sentimental attachment seems more likely

to arise, so such moral concerns would extend more broadly. Suppose, though I think it implausible, sentimental attachment forms between a driver and its avatar quite generally. Still, the commonplace nature of sentimental attachment does not give us the distinctive moral concern that I seek. In this case there does not seem to be anything that latches into how avatars change our lives in the world; rather, they are just like any other object that someone might develop a sentimental attachment to, like favorite dolls and fossils.

With either kind of psychological attachment as a ground for special moral status for avatars, the moral status of avatars depends on whether certain psychological connections exist. So, whether it is morally permissible to, for instance, virtually distort, beat, rape, kill, or delete one avatar but not another *hinges* on how its driver feels about it or is psychologically connected to it. Maybe all that matters morally about avatars lies in our psychological connections to them as described.

But these connections fail to offer us compelling reasons to give avatars special moral status that satisfies the constraints above. At best, the connection that ties avatars tightly to self-identity like a ventriloquist who strongly self-identifies with his dummy applies only in extraordinary circumstances and so fails to satisfy our generality requirement, while sentimental connections that place avatars in the toy box with favorite dolls applies in pretty mundane circumstances and so fails to latch onto the distinctive role of avatars in our lives. So, even if we accept the psychological thesis, on neither approach does it suffice for us to meet the challenge at hand.

Part 2: The positive project

The shape of my positive argument for the claim that avatars have special moral status is straightforward. I take as given that among our moral rights is a stringent bodily control right. We have a stringent bodily control right because human bodies have special moral status. What I seek to show is that *the reasons we have to confer special moral status on human bodies also serve as reasons to confer special moral status on avatars*. Therefore, avatars have special moral status relevantly similar to the moral status of human bodies. And so, for the sake of consistency, we should endorse avatar control rights. This is an argument by analogy. I find its consequences appealing for concerns that in part trace back to the worries about virtual rape with which we began.

We may appeal directly to bodily control rights to explain the moral impermissibility of physical rape. Likewise, perhaps we may appeal directly to avatar control rights to explain the moral impermissibility of virtual rape. Given stringent control rights, the moral wrongness of the violation does not depend upon the victim's psychological state. A rapist may not violate the bodily control rights of a victim because she is too drunk to care; a virtual rapist may not violate the avatar control rights of a victim because the driver is not properly psychologically invested in the avatar. Moreover, in neither case do we let the offender of the hook because the victim is in the wrong place. We do not excuse the physical rapist because the victim walked alone in the park; we do not excuse the virtual rapist because a newbie entered Second Life sans chaperon. Finally, the moral impermissibility of rape remains regardless of how hard the victim fought. Physical rape remains a violation even if the victim does not fight back; virtual rape remains a violation even if the victim did not turn off the computer. For, what matters with respect to physical or virtual rape is that a person's stringent bodily or avatar control rights were violated.

If my argument succeeds in supporting a stringent right to avatar integrity, then we have a solution to the challenge at hand. The argument applies quite generally to avatars, it draws on the distinctive role of avatars in our lives, and it appeals to reasoning that we already accept as demonstrating special moral status for bodies.[21]

But why should we go along with the key claim that reasons that motivate special moral concern with human bodies likewise motivate special moral concern with avatars?

I want to set to the side one possible answer to this question. We might try to say that we have reasons that motivate moral concerns with both avatars and bodies simply because avatars are part of our bodies in the relevant moral or material sense of "body." In this case, the outlines of organic human bodies would not match the outlines of the bodies over which we have stringent bodily control rights. Rather, on such an approach, avatars would be extensions of bodies on the proper construal of bodies. Our bodies would in a sense be cyborgs.

I find it plausible that the material or moral contours of our bodies should include things that function as parts of us in important ways even if they aren't part of our original equipment.[22] Such retrofitted parts might include transplanted hands, hearts, and faces, installed manufactured corneas, mind-controlled bionic arms, cochlear implants necessary for hearing, or Harbisson's eyeborg that allows for an extraordinary sensory capacity. If we expand our understanding of body to include such things, then it may be also reasonable to expand that understanding to reach into the virtual world to include our avatars. On such a view, it would follow that avatar control rights would be part of our bodily control rights.

The way we understand our embodied selves may well need to change in light of technology. To exclude various retrofitted parts may well be morally arbitrary. For it does not seem to me that the moral value of my body lies with it including only my original parts, or only meat parts, or only irreplaceable parts, or only parts that deliver nothing above normal human functioning. But I will not pursue this expanded body approach here because it introduces a difficulty that we may bypass.

A coherent argument for the claim that avatars should be parts of our expanded bodies requires a principled non-biological criterion to decide when something properly counts as a part of an expanded body and when it does not. We would then show that avatars should count as part of our bodies based on such a criterion. However, to identify such a criterion is a challenging task.

If we do not appeal to biology, do we instead appeal to some other material criterion, or must it be a moral criterion? At least at first glance, it would need to be both. A criterion that provides no material guidance on what might be included in my body seems to give up too easily that we are embodied material beings. Yet, plausibly the criterion would also require some moral component.[23] (This is why I hedged above with my use of the "material or moral" locution.) For instance, suppose we decide to extend our bodies to include objects that in a meaningful sense directly connect to our sensory capacities, such as cochlear implants or replacement corneas. We face the normative question of whether we should include only those parts necessary for something like normal human functioning or also extend our bodies to incorporate new capacities, like the eyeborg. We can identify devices that alter our sensory capacities with material criteria, but it is a normative question which of these would properly belong to our extended bodies and why we should care about those capacities in the first place. To respond brings us into debates about the ethics of human enhancement, and so on. Fortunately, we may avoid the claim that avatars are parts of our moral or material bodies to say they have special moral status, so we need not try to establish a criterion to define non-biological moral or material contours of bodies. Thus, let us leave open the question of whether we

should change the material or moral contours of our bodies to include avatars as parts.

Instead, we will pursue the weaker claim that avatars are *like* parts of our bodies in important ways, namely, in ways similar to those that make us care so much about our bodies in the first place. If we successfully demonstrate similar reasons motivate moral concerns about both bodies and avatars but not dolls or puppets, then we have good reasons to believe that avatars are not mere dolls or puppets.

Let us now proceed to see why avatars are like parts of us in morally relevant ways. The idea is that bodies have special moral status because they do certain things for us. Avatars also do similar things for us, so they are like bodies in those ways. Now we must ask: What are those certain things? What gives bodies their moral oomph?

It is straightforward to compose an uncontroversial partial list of what makes our bodies matter. Without our bodies, we cease to exist in the non-virtual world. Beyond mere existence, our bodies give us access to and causal power in the world. Bodies are critical to our autonomy and agency: we cannot act in the world except through the use of our bodies. We need our bodies to perceive and apprehend the world. Our bodies provide the pathways through which we sense pain and pleasure. Our bodies allow us to function as social beings. I suggest that we can compose a similar list for why avatars matter morally, as long as we allow that the virtual world is real.

Because we have decided not to let our argument depend upon expanding our body to include virtual parts, it does not make sense at this juncture to say that we (or parts of us) would cease to exist in the virtual world without our avatars. For to say that they are part of our existence is to make them parts of us again. But, I think we may say with little controversy that without our avatars, we cease to have any possibility of a presence in the virtual world. This is not merely 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, genuine presence involves a robust sense of being on the scene in some way. Beth Coleman's rich discussions on presence are helpful for those who wish to cash out this idea.[24] Dolls and puppets do not offer us existence or presence anywhere, regardless of how we may use them to express ourselves. Beyond presence, our bodies give us access to and causal power in the world.

For instance, our avatars are critical to our autonomy and agency: we cannot act in virtual reality except through the use of our avatars. A denizen of Second Life cannot build a virtual house or try a virtual hairstyle or speak from a virtual soapbox without acting through her avatar. Actions taking place solely in the non-virtual world or in our purely mental constructions have no impact on virtual reality. Nothing about an absence of dolls or puppets constrains our power in the non-virtual world. A ventriloquist deeply attached to a dummy may find herself less able to act, but such a result would be due to a psychological barrier, not due to the absence of her dummy itself.

We need our avatars to perceive and apprehend the virtual world. As a practical matter, our avatars serve as our eyes and ears in virtual reality. We cannot, for instance, peer into the space of Second Life without them. Likewise, it is a practical matter that our eyes and ears serve as our sensory organs in non-virtual reality. They are already partially replaceable in the form of cochlear implants and artificial eyes. The contingency between avatar/body and our sensory powers does not erase their importance to us.

Our avatars also give us a way to apprehend the world. We cannot fully appreciate what it is to participate in virtual reality without doing so through our avatars. Dibbell recognized this early on when he noted his shift in how he understood the meaningfulness of participation in LamdaMOO.[25] I imagine the same is true of travel not across the virtual border but through our atmosphere to space. I can read all I like, use simulators extensively, but unless I have the chance to peer at the earth from the moon or cavort weightless in space (safely tethered to the space station, of course), I don't think I can fully apprehend what it is like to experience life in space. In the first case, we must drive an avatar in virtual reality. In the second case, we must envelop our bodies with protective gear to leave our atmosphere. Perhaps it is worrisome that protective gear might sometimes get a bit of extra moral status, perhaps not. Either way, dolls and puppets do not augment our ability to apprehend the world in similar ways.

Arguably, our avatars also provide the pathways through which we sense virtual pain and virtual pleasure. For instance, on a pleasant note, Coleman shares her non-virtual world experience of physical memory of soaring through the skylight in the MIT dome in virtual reality. Her pleasurable physical sensation traces directly back to her avatar experiences.[26]

Finally, our avatars certainly allow us to function as social beings in virtual reality. Individuals have rich and meaningful social lives in virtual space. Genuine friendships form, love develops. These social interactions and relationships are not pretense, but important parts of our lives as social beings that often affect our non-virtual lives and vice versa. In itself, nothing about an absence of dolls or puppets constrains our ability to be social creatures. A ventriloquist deeply attached to a dummy may find herself less able to socialize, but such a result would be due to a psychological barrier, not due to the absence of her dummy itself.

I have not aimed at an exhaustive list for how bodies and avatars matter morally. Nevertheless, I think the important and unique role that avatars play in our lives in ways so analogous to the important and unique role that bodies play in our lives demonstrate that it is not untenable to think that avatars may have special moral status of a kind and strength that in general does not attach to mere property, dolls, or puppets.

My position is that the sort of moral concerns considered above explain why our bodies have special moral status that justifies stringent bodily control rights. Intuitively: My body, my choice! Roughly, what do our bodily control rights look like? There is broad and deep consensus that we have stringent bodily control rights. Among other things, such a right protects against things like physical coercion, unwanted physical contact, invasive contact without consent, harm, or damage to one's body, medical procedures without consent, and use of one's body without consent. The burden of proof to justify an exception to the right is very high and usually includes an appeal to consent or to protection of oneself or others.

Similarly, I take it the sort of moral concerns considered above suggest that avatars have special moral status deserving of stringent avatar control rights. Intuitively: My avatar, my choice! Roughly, what might avatar control rights look like? Such a right might protect against: deletion of the avatar without consent, use of one's avatar without consent, forced virtual relocation of the avatar, any harm or damage to one's avatar without consent, and any change of the avatar's code without consent. And, of course, it would protect against virtual rape.

If I make my case, the response here to the original challenge succeeds. The burden of proof to justify an exception to the right is plausibly higher than it is for mere property for reasons analogous to those for why bodies matter more than mere property. The special moral status applies to avatars quite generally—it involves no requirement for drivers to have extraordinary or sentimental attachment to artifacts. And the

special moral status traces directly to the distinctive role that avatars play in our virtual lives.

Objections

We have considered an argument for stringent avatar control rights based on the special moral status of avatars. In closing, let me defend this view against four objections.

First, one might object that avatar control rights are no more than a particular subset of property rights. Some, for instance, hold that bodily control rights are no more than a particular subset of property rights, they just happen to be very stringent control rights.[27] So, we really have no reason to see avatar control rights as falling outside the scope of property rights. I say fair enough. If bodily control rights are a stringent set of property rights, we have no reason to think avatar control rights would be different in kind. But, this does not affect my position. I take no stand here on the relationship that obtains between personal property rights and bodily control rights. Bodily control rights may be a special kind of personal property right or not. What matters in my remarks above is that the stringency of bodily control rights hinges on human bodies having greater moral status that that of *mere* property. My aim has been to make the analogous claim that avatar control rights may likewise hinge on avatars having greater moral status than that of mere property.

Second, one might deny that human bodies have special moral status in the first place. An anonymous reviewer articulated this challenge quite nicely:

> Granted that the worst thing that can befall us is the destruction of our bodies, since this obliterates all autonomy, agency, thought, and our very being, and that consequently the destruction of our bodies is a greater harm than the destruction of any of our possessions, it does not follow that bodies have any "special" moral status. Even if the worst harm that can befall us, viz. death, is a bodily harm, it does not follow that every violation of bodily integrity is a greater harm than any violation of our property rights or other harm. Forcibly cutting off my hair, a violation of my bodily integrity, is surely a less serious harm than . . . destroying a manuscript to which I've devoted years of work. The received view is, indeed, that bodies have some "special" moral status, but I have yet to see a convincing defense of this doctrine.

One small quibble: to say that bodies have special moral status does not mean that moral concerns that trace to that moral status may never be overridden by other concerns.

Nevertheless, I have real sympathy with the meat of this criticism. Consider, for instance the case of Jan Scheuermann, a quadriplegic who can now feed herself thanks to a brain-controlled bionic arm.[28] It strikes me as ludicrous to think that my fingernails have greater moral status than Scheuermann's bionic arm simply because they belong to my natural body. Not unreasonably, we might use my position above as a reductio against the special moral status of human bodies. For once we see that even avatars may share the features in virtue of which we care so much about our bodies, mapping special moral concerns onto a contiguous hunk of organic matter looks arbitrary.

Perhaps we should scrap the idea of *human* bodies as the bearers of special moral status. But, if so, it does not follow that we should give up on the notion of the special importance of our bodies (in some sense) altogether. Rather, it suggests to me that we should reconceive of the moral or material contours of our bodies to better capture those things in the world, virtual or non-virtual, that play the sort of important role in our lives that I've described in my discussion above.

Fortunately, my position in this paper does not require us to complete that more difficult project. If it turns out that my argument by analogy works because, once properly drawn, the contours of our bodies include avatars, so much the better.

Third, one might accept the trauma experienced by victims of virtual rape as genuine; yet reject the significance or appropriateness of the reaction. On this note, an anonymous reviewer writes: "A century ago when movies were a novelty audiences ducked when trains rushed at them on the big screen. *Should they have?* These virtual trains couldn't really harm them and, arguably, neither can attacks on avatars really harm their drivers—even if they can, as it were, *virtually harm* them."

On my understanding, the thought is that the trauma of virtual rape, though real, arises from habitual psychological responses that confuse real and pretend threats. When we become familiar with the new stimuli, the problem goes away. So, once we understand how movies work, we *should not* duck to avoid a movie train. Perhaps, once we understand how virtual reality works, we should not respond to virtual rape as we would physical rape. In some respects, this is true. We should not duck at movie trains and we should not go to the hospital to complete a rape kit after a virtual rape. But the root idea that ducking movie trains and experiencing trauma from virtual rape both amount to inappropriate responses to pretend stimuli does not ring true to me.

There are important differences in the cases. Consider, for instance, that we do not ordinarily use the trains we see on the movie screen to do anything in the non-virtual world, whereas we regularly use avatars to do all sorts of things in the virtual world. I may leave a cinema to avoid scary train images and preserve my ability to perceive, apprehend, and interact with the non-virtual world. But I diminish my ability to perceive, apprehend, and interact with the virtual if I forfeit my avatar to avoid a virtual assault. Movies are not interactive, virtual reality is. Movies play to a general audience; the images on them do not change based on an individual's response or fear. In contrast, the driver of a virtual rapist targets an individual and presumably makes choices in light of the virtual victim's responses that the driver of the virtual victim controls. We may avoid movies altogether yet still have plenty of space to develop and pursue our own conceptions of a good life. Realistically, most people cannot dispense with their presence in virtual reality without significantly limiting their choices.[29]

The objector would deny such differences to assimilate virtual rape of an avatar to something like the pretend rape of a doll. But what happens if we make a non-virtual doll more like an avatar in the virtual world?

> **DummyGodWorld:** A dummy god makes our abilities to experience and act in the world dependent upon ventriloquist dummies in peculiar ways. To perceive, to apprehend, and to act in this world, everyone must be a dummy driver. Consider what life is like for someone in this world. To open her eyes, a person must also open the eyes of her dummy. To speak to anyone, she must throw her voice and speak through her dummy. To walk down the street, she must carry along the dummy, swinging it in such a way that it appears to be walking. To work, to go to school, to dance, she must at the same time drive her dummy to engage in similar activities. Though no one confuses her as a person with her dummy, or believes that the dummy acts independently of her, people recognize her by her dummy and address her dummy. While she is the person who experiences, perceives, and acts in the world, her ability to do so is contingent on her ability to drive her dummy.

In **DummyGodWorld**, if one person drives its dummy to assault the dummy of another person, what do we think? The assaulted dummy has its clothing torn, its orifices penetrated. The attacking driver narrates the brutality via the attacking dummy. The victim driver could, of course, simply drop her dummy to in some sense avoid the attack, but would thereby at least temporarily lose her ability to experience and act in the world. Would a person traumatized by an attack on her dummy be making a mistake like the cinematic newbie who ducks a movie train? I suggest she would not. Of course, the dummy assault would not be like a physical assault. But, because of the weird contingency between driving dummies and ability to experience and act in the world, the harm retains significance beyond a pretend attack on a mere doll. We should not be expected to habituate ourselves to simulated dummy assaults in order to function in the non-virtual world of the dummy god. Likewise, a call for habituation to simulated assaults in virtual worlds is not the appropriate moral response to virtual rape. For, there exists the same weird contingency between driving avatars and our ability experience and act in virtual space.

Fourth, one might argue that I have gone too far with respect to the moral significance of avatars. As I said at the outset, virtual rape is *not* morally on a par with physical rape. By arguing for the moral status of avatars by way of the moral status of bodies, we seem to have undermined our grounds for seeing physical rape as the greater moral violation.

Here I have two replies. First, if we take the path I set aside and count avatars as parts of our bodies, we still need not accept the consequence that the moral distinction between virtual and physical rape collapses. Even for our physical bodies, we distinguish between violations that involve parts of our bodies in various ways. For instance, it is worse to cut off my pinky than my hair without my consent.

Second, if we take the path of arguing for the moral status of avatars via an analogy with the moral status of bodies, we need not take that analogy too far. Overall, the importance of our bodies trumps the importance of our avatars when it comes to their role in agency, perception, apprehension, pain, pleasure, love, and friendship. To say that similar moral reasoning elevates both the moral status of avatars and bodies above puppets and dolls is not to say that it elevates both to the same level. Indeed, it strikes me as right to say that avatars fall between bodies and mere property in terms of their moral weightiness.

Such a conclusion gives us a way to make sense of the following insight shared by Michael Bugeja in his thoughtful discussion on avatar rape, even if we resist the word "imaginary":

> Tim Guest, author of *Second Lives*, believes sexual assault in virtual worlds is real and imaginary. "As the saying goes, the thought is written in water, and the deed is written in stone. Events that take place in virtual worlds seem to lie somewhere in between, a kind of water with memory."[30]

Acknowledgements

Thanks to Keith Miller, Mary Sheila Tracy, and the excellent students in their spring 2012 RoboEthics course as well as the participants of the fall 2012 UIS Philosophy Faculty Seminar for comments on earlier versions of this project. Thanks also to a particularly helpful anonymous reviewer for useful comments and criticisms.

Notes

1. Dibbell, "Rape in Cyberspace."
2. Here, I specifically exclude the phenomenon of games in which players consent to participating in rape scenarios.
3. I mean this as a general moral claim rather than a universal moral claim that allows no exceptions. No doubt, the contray philosophers among us, myself included, may come up with a case in which the opposite is true.
4. I take it that the possibility of virtual property was more controversial when it seemed that the virtual world was pretend, disconnected from the real world, rather than a part of it. But now I think we must agree with it. Certainly we may have property rights over virtual property, and avatars may be among our possessions.
5. From Bugeja, "Avatar Rape": "Second Life advocates often note that avatar assault is easily avoided; you can teleport away, they say. Linden Lab recommends muting voice during verbal assaults. 'Click! Problem solved', it states."
6. I think for any object we might make up a story on which it would make sense to say the object would have special moral status in some sense. Philosophers are clever like that.
7. I say "in standard cases" to allow for exceptions in unusual circumstances.
8. I do not want to take the position that we can find no similar issues arising throughout the ages, well in advance of today's marvels of technology. The issue has just become more pressing through its new prevalence.
9. Harbisson, "I Listen to Color."
10. The reader may be less familiar with *Surrogates* than *Avatar* (both films from 2009). Here is a brief description of *Surrogates* from IMDB: "Set in a futuristic world where humans live in isolation and interact through surrogate robots, a cop is forced to leave his home for the first time in years in order to investigate the murders of others' surrogates." http://www.imdb.com/title/tt0986263/.
11. For an interesting discussion of the porousness between the virtual and non-virtual world and how they may fit together, see Coleman, *Hello Avatar*, especially "Chapter 5: X Reality: A Conclusion."
12. For discussions on moral problems in virtual reality, see for instance, Huff, Johnson, and Miller,"Virtual Harms"; Powers, "Real Wrongs"; and Wolfendale, "My Avatar, My Self."
13. I mean these as *prima facie* moral rules of thumb that may allow for exceptions.
14. See Coleman, *Hello Avatar*, chapter 1: "What is an Avatar?," in which she offers a very helpful discussion on how our understanding of the metaphysics of avatars is changing as our virtual lives expand both online and in their connections to our non-virtual lives.
15. A useful definition from Coleman's glossary in *Hello Avatar*, p. 187: "Avatar Incarnation of a deity in mortal form, often as a hero (Hindu); a computer-generated figure animated by player or participated in online media context, such as a virtual world; the gestalt of images, text, and multimedia that facilitate presence in networked media."
16. As representative of such an approach see, for instance, Wolfendale, "My Avatar, My Self." At times, Wolfendale appears to argue for the stronger version of the thesis, at times the weaker thesis.
17. Wolfendale, "My Avatar, My Self."
18. O'Connell, "Right Word, for Dummies."
19. Ibid.
20. Whitty, "'Dumbstruck' Review."
21. I say "reasoning that we already accept as demonstrating special moral status for bodies," but I would more accurately say "reasoning that many have accepted as demonstrating special moral status for bodies." There are deep complications here that I sought to set to the side for the purposes of this paper. But, an anonymous referee rightly urged me to clarify my argument on this point, which I take up in my discussion on objections.
22. Of course, the idea of original body parts involves issues about persistence of bodies over time because of material changes in our bodies. But let us not get sidetracked by this problem here.

23. Compare, for instance, similar concerns that arise when we seek to make sense of human nature in a way that connects to morality regardless of whether we involve add-on virtual or non-virtual parts.
24. See Coleman, *Hello Avatar*, chapter 4: "Presence."
25. Dibbell, "Rape in Cyberspace."
26. Coleman, *Hello Avatar*, 47.
27. Thanks to William Kline for challenging me on this point in an early discussion.
28. Fox, "Woman Uses Thought Control."
29. I aim this also as a response to an anonymous reviewer's question about whether the harm of a virtual rape crosses the virtual border to cause real harm to the avatar's driver.
30. Bujega, "Avatar Rape."

Bibliography

Bugeja, M. "Avatar Rape." *Inside Higher Ed*, February 25, 2010, http://www.insidehighered.com/views/2010/02/25/bugeja#ixzz2EX3PjBWt.

Coleman, B. *Hello Avatar: Rise of the Networked Generation*. Cambridge: MIT Press, 2011.

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.

Fox, M. 2012. "Woman Uses Thought Control to Eat Chocolate." *NBCNews.com*, December 17, 2012, http://vitals.nbcnews.com/_news/2012/12/17/15955022-woman-uses-thought-control-to-eat-chocolate?lite.

Harbisson, N. "I Listen to Color." *CNN* website, September 10, 2012, http://www.cnn.com/2012/09/09/opinion/harbisson-hear-colors/index.html.

Huff, C., D. G. Johnson, and K. Miller. "Virtual Harms and Real Responsibility." *Technology and Society Magazine, IEEE* 22, no. 2 (2003): 12–19.

O'Connell, S. "The Right Word, for Dummies." *Washington Post*, April 22, 2011, http://www.washingtonpost.com/gog/movies/dumbstruck,1207292/critic-review.html.

Powers, T. M. "Real Wrongs in Virtual Communities." *Ethics and Information Technology* 5, no. 4 (2003): 191–98.

Whitty, S. "'Dumbstruck' Review: Documentary Captures Drama of Ventriloquism." *The Star Ledger*, April 22, 2011, http://www.nj.com/entertainment/movies/index.ssf/2011/04/dumbstruck_review_documentary_captures_drama_of_ventriloquism.html.

Wolfendale, J. "My Avatar, My Self: Virtual Harm and Attachment." *Ethics and Information Technology* 9, no. 2 (2007): 111–19.

A Philosophy of the Web

Sidey Myoo
Jagiellonian University, Krakow

> Computers, too, lead us to construct things in new ways. With computers we can simulate nature in a program or leave nature aside and build second natures limited only by our powers of imagination and abstraction. The objects on the screen have no simple physical referent. In this sense, life on the screen is without origins and foundation. It is a place where signs taken for reality may substitute for the real. Its aesthetic has to do with manipulation and recombination.
>
> – Sherry Turkle, *Life on the screen. Identity In the Age of Internet*

> What I'm saying is that you have to think about technology, you have to use it, because in the end it is in your blood. Technology will move in and speak through you, like it or not. Best not to ignore.
>
> – Tim Etchells, *Certain Fragments: Contemporary Performance and Forced Entertainment*

Introduction

My aim is to propose broad outlines for a philosophy of the web, after which I will provide a description of the Academia Electronica, a type of online university, as an illustration of how the philosophical ideas outlined may be applied on a practical level. In particular, I will describe the Academia Electronica in the context of an ontological postulation with respect to electronic reality regarded as a sphere of being. What underlies my choice of subject matter is philosophical reflection on human engagement with the web, the type of activity engaged in, and the time spent on the web. The main idea is that the creation of one's own personal space on the web and the intensity of communication mean that these phenomena cannot be considered unreal and therefore cannot be described as, for example, artificial.

I decided on this dual approach for the article since I recognize that theoretical analysis is not necessarily adequate or convincing if it fails to entail real consequences. A philosophy of the web requires statements concerning genuine human activity in the electronic reality as well as ontological propositions on this sphere of being, unless someone who spends hours each day working at a computer insists that they are engaged in something artificial or unreal, or an imitation of reality.

A story

Late one night, deeply immersed in the electronic world of Second Life, I made my way to the coast to spend a few moments on the beach before leaving the online world. On the interface I switched the time to sunset and sat back in a deckchair. After a while, a woman (from Holland, as I recall) came up to me and after a short, customary greeting told me that she had just become homeless. At first I was puzzled as to how someone sitting at a computer could be homeless; however, it turned out that the home in question was not a physical one but one she had built with her Second Life husband and had been living in for the previous nine months. Quite apart from her relationship in Second Life, where she spent time every day and had a child (a bot), she was married in the physical world. On the day I met her, her Second Life relationship had come to an end and she had, in that sense, become homeless.

A philosophy of the web

This story goes some way to explaining my understanding of the philosophical issues relating to the ontology of electronic being and the human person. In general terms, this subject matter arises out of the fact that humans are increasingly active on the web, both quantitatively and qualitatively; their attention is drawn to it and they find ever more content within it, frequently at the expense of certain forms of being in the physical world. This is not to say that I regard online phenomena merely as modes of communication utilitarian in character, but rather as having the nature of real human activity with an ontological and anthropological dimension. Since I regard online events as being just as important, authentic, and real as those in the physical world, I use concepts of both electronic and physical reality. I attach particular importance to worlds created by means of interactive 3D graphics; thus, what I have in mind when I refer to "the electronic world" is an electronic reality in which people gather for no other reason than to participate

and where one's motives for spending time are existential. I emphasize the significance of electronic worlds because they have the power to assimilate various kinds of human activity, at which point they become of interest to the philosopher.

> By "world," I mean just that: a place with earth and sky, rivers and trees, deserts and meadows, with a sun that makes regular trips across the horizon, a moon that crests in its passing, and most of all the natural elements that quietly comprise a place: gravity and wind, and an ocean that responds to both. All of this, depicted on the computer monitor before you. There are people there, too.[1]

I propose a philosophy of the web as a branch of philosophy combining all its various strands. Furthermore, I acknowledge that analysis of online phenomena is fundamental to understanding humankind and the world in the present day, and that there is indeed an onus on philosophy to explain these phenomena at a more fundamental level than would be the case for other fields of learning. This involves directing philosophical analysis to electronic reality and the technology with which it is so thoroughly imbued, which means understanding philosophy as the science of technological imponderables. I would argue that the analysis and constant re-evaluation of these phenomena is incumbent on the humanities in their relationship with humanity, which, finding itself in an ever more intimate coexistence with technology, does not necessarily perceive the profound changes taking place, seeing their invasiveness and paradigmaticality as utility. We work, spend time, and have emotional experiences in both realities, and the activity and meaning which we bring into effect and to which we are subject are equally real to us in either: people and things are capable of authentic and real existence in both spheres.

Technology is taking over many aspects of human life, enveloping us in a constantly expanding, complementary sphere, both in the electronic reality of the web and in the physical world in the form of smart appliances. It is, I believe, illusory to regard computers merely as tools, and even erroneous since the computer (the interface) is actually the gateway to another world:

> Thus, image becomes image-interface. In this role it functions as a portal into another world, like an icon in the Middle Ages or a mirror in modern literature and cinema. Rather than staying on its surface, we expect to go "into" the image.[2]

The computer may be seen more abstractly as an evolving device, constantly enhancing its possibilities. From this perspective, a personal computer becomes not just a concrete object but the manifestation of a technology at a particular stage of its development. This involves understanding technology as developing at an incomparably faster pace than other fields, giving rise to futurological extrapolations that further development of technology will bring in its wake manifest consequences for humanity.[3] Statements such as: *In the future, processors and computers will be faster*, should be regarded as reasonable extrapolations and not casual opinions of no scientific consequence. In evolutionary history, technological development goes on unabated, and with each technological advance, humanity is systematically alienated from the natural world, which itself is being transformed into an artifact by the power of technology.

The fundamental philosophical questions to be resolved are: What constitutes reality in the contemporary world? How should philosophy relate to various kinds of reality: physical, electronic, immaterial, or a hybrid form of being? What is the value of gaining access to the web through various devices, subsequently spending hours on end online, and benefiting from continuity of communication? Which ontological categories should be assigned to immaterial, electronic forms of being that have their origins in the physical world, and how should their value be assessed?

Let us leave behind the bimodal reality as described in two different ontologies. I do not intend to discuss issues of augmentalism and immersionism, although I have reservations about augmentalism precisely because it has its roots in two different ontologies. Immersionism, on the other hand, assumes the transfer of intentionality from the physical to the electronic world while partially inhibiting certain types of activity, for example, sensory, in the latter. I am not sure how to understand the concept of extending the boundaries of the physical world when it might rather be a case of adding the ontologically different electronic space to the physical world. Buechner's remarks concerning the alterity between physical and electronic being in the context of augmentalism are intuitive:

> The claim of this paper is that one kind of augmented reality is philosophically incoherent. That is, there are a priori reasons to believe that it cannot happen. It is not that the concept makes sense, but is either physically or technologically unachievable. Rather, the very concept is incoherent. It is metaphysically impossible. [. . .] My claim is not about limitations of the physical world, but rather about the concept of reality augmentation and the metaphysical limitations imposed by a philosophical theory of fictional entities.[4]

Intuitive propositions that web-based phenomena may be regarded as a type of reality began to appear in the literature in the 1970s. Although they grasped the sense of reality arising from the development of electronics, they were still deeply rooted in the idea of electronic reality being something artificial and unreal. This approach doubtless arose from a deep-seated and ontologically weak understanding of virtual reality as something unsubstantial and ephemeral rather than as a distinct and ontologically convincing form of being, an example of which is someone claiming that he possesses virtual money. This might, for example, mean that the individual in question has been promised money, which, not yet being in their possession, is, in a certain sense, non-existent; or it may refer to money which the individual really possesses in their account but which exists electronically rather than physically. Thus, an ontological distinction can be made between virtual and electronic: virtual money is not real and is not held in an account but which is, for example, expected; money held electronically rather than physically in an account, on the other hand, is real rather than virtual.

Historical, yet significant on account of her position on the ontology of electronic forms of being, are the findings of Sherry Turkle, which stem largely from psychological analysis:

> What is real? That question may take many forms. What are we willing to count as real? What do our models allow us to see as real? To what degree are we willing to take simulations for reality? How do we keep a sense that there is a reality distinct from simulation? Would that sense be itself an illusion?[5]

In an ontological sense, the emergence of virtual reality can be traced to Myron Krueger in the early 1970s. However, in common with other contemporary views (*virtual realism*, Michael Heim; *virtual realm*, Margaret Morse; *new nature of reality*, Nicole Stenger; *parallel universe*, Michael Benedikt; *cyber world*, Hans Moravec; *work space*, Steve Pruitt; *computer culture*, Dave Healy; *virtual community*, Howard Rheingold), he failed to grasp the full ontological meaning which would have enabled

him to understand electronic reality as a sphere of being. Moreover, Krueger also made use of another idea, *artificial reality*, which actually militated against ontological analysis since it introduced the notion of artificiality. Nevertheless, in my estimation Myron Krueger came closer to seeing virtuality as reality in the ontological sense than any other theoreticians I am familiar with who also entertained intuitions in this area. What he lacked, in my opinion, was philosophical analysis, which would have placed his intuitions on electronic reality on the level of philosophical categories.[6]

Once we accept the realness of electronic reality, apart from fundamental ontological issues, there are also anthropological ramifications. A person can simply become addicted to communicating or being in electronic reality. Rejecting this situation may give rise to technological exclusion. Giving up using a cell phone or emails can lead to an existential bubble in which one is soon faced with limitations such as the inability to communicate with others: "The concern has been that if we are spending more time in virtual rather than in face-to-face communication, our weak ties may grow but strong ties shrink."[7]

The expanding sphere of electronic being makes various kinds of online experience possible. I regard such experiences as the natural development of events in the life of the modern individual: by this I mean a philosophical interpretation connected with, for example, self-creation; in other words, the emergence of an online identity or the existence of knowledge as transcendent in the shared electronic sphere of being.[8]

When an individual enters the online world, particularly a graphic 3D environment, one discovers a space for self-creation and can begin to effect changes which also extend to one's physical existence. And here lies the essential point: an individual can exist in electronic reality in new and diverse ways. This is the difference between instant messaging and the electronic world. Depending on technological possibilities, an individual can introduce various content online, the most important element of which is one's emotions. There are also axiological implications: values, which may be present in any form of human reality, have the same meaning as when originally encountered in the physical world. Thus, in electronic reality an individual can find a world of feelings and spiritual values, such as truth and falsehood, which are not established by a particular kind of reality but by an individual's activity within it. The value system that an individual encounters in electronic reality can also suggest choices to which they will be subject. Two questions arise from this. The first relates to the fact that an individual can see on the screen what they are saying or how they appear in the form and behavior of their avatar; this may lead them to modify their actions and learn from the experience since it is crucial for them to exist in electronic reality. In order to maintain positive activity within a nonlinear structure of contacts, such behavior is necessary to eliminate negative values. This is quite distinctive since technology creates situations that cannot exist in the physical world, which is mainly due to nonlinearity and the ease with which emotions can be expressed online. The second question has to do with situations where an individual easily becomes emotional, which may or may not give them a sense of the significance of the value experienced and of responsibility for their behavior. This is a rather common experience in electronic reality when an individual is faced with choices straddling the physical and electronic worlds. If an existential balance between the two worlds is not maintained, someone who has important issues in the physical world can easily be faced with similar issues in the electronic world and will have to make choices between the two.

> I argue that *it is in being virtual that we are human*. Virtual worlds reconfigure selfhood and sociality, but this is only possible because they rework the virtuality that characterizes human being in the actual world.[9]

This involves both the psychic and the corporeal:

> A virtual being has mystery—that of the coevolution of man and machine, that of the redefinition of the body, of the organic, and of evolution. A virtual being is a perception that is alive.[10]

An important factor is self-expression, which in the 3D world begins with the physical appearance of the avatar. The individual goes about the electronic world in a form they have created themselves and with a name they have given themselves, capable of making friends or falling in love. Being in the electronic world becomes very pleasurable when, as a matter of course, one can access a space where all the problems of the physical world have been removed. Continual participation in an environment like Second Life can turn into an authentic existence.

The moment we accept the electronic environment as a sphere of being and an alternative reality to the physical world, the events and experiences of the physical world can take on a credible and valuable form in electronic reality. Things really happen, but just differently from in the physical world because they are governed by a different ontology. Unless the electronic sphere is recognized as a type of human reality in its own right, each activity in the electronic reality will to some extent be seen as separated from the physical world but complementary to it; it will never attain its fullness but remain a hybrid whose essence is to be found in the physical world. The fundamental point here is one's philosophical attitude to the world. Without, at this stage, going into the possibility of affirming the existence of any particular reality, the individual lives with some sort of conviction about the existence of the physical world and has no need to cogitate on the matter: that is for philosophers. A similar conviction is also in evidence when I affirm the existence of electronic reality. It is something I accept and seek to substantiate as a developing sphere of being, doing so from the perspective of a philosopher living in the contemporary world.

> For the past three decades, I have been fascinated with the construction of identity and how it affects culture: the symbiotic relationship between the real and the virtual, and how identity reacts and shifts when processed through manipulated time.[11]

One day I realized that what I was doing in electronic reality amounted to genuine engagement. This intuition led to the setting up of online university courses in the form of Academia Electronica in Second Life.[12]

The idea of academism

In this section of the article I would like to illustrate the practical dimension of a philosophy of the web. To this end, I will describe the Academia Electronica, a non-institutional university in Second Life in which I have run official, academic lectures for five years. Apart from lecture courses, individual lectures are also given by invited guests as well as undergraduate and postgraduate students. Most of the lectures are archived in the form of audio recordings on the academy's website. The Academia Electronica embodies the idea of academism in that it extends and diversifies the content of academic life possible in e-learning.

The academy is mainly concerned with examining the multifarious issues that arise when the electronic environment is regarded as a sphere of human reality. It describes electronic

reality from the perspectives of philosophy, cultural studies, sociology, psychology, and other disciplines. While the academy provides a platform for discussing philosophy, it itself is a subject of philosophical enquiry and a laboratory of the humanities. It asks whether electronic being can really exist and an online identity really be created, and whether values can exist in a nonlinear system of human communication.

I chose Second Life firstly because I realized that it is the best form of electronic reality: an electronic world in which various aspects of academic life can be present; and secondly because I am convinced that technological development will affect the quality and length of online participation, especially in electronic worlds, leading to ever more widespread avatarization. Avatarization indicates a state of affairs that enables individuals, in the form of their avatar, to engage in unrestricted activity in electronic reality (including professionally), to maintain other contacts, and to possess goods. I also realized that Second Life is the best method of academic contact since it not only enables communication with students but, by its very nature, allows the expression and exchange of views. For example, part of my contact with Masters and PhD students is through Second Life. This sometimes takes place in the evening, often around a campfire. I believe such conversations can be more effective than institutional meetings in a physical university where the environment itself (being in the professor's study and being faced with the barrier of the professor's desk) determines the nature of contact, potentially inhibiting the student from engaging in philosophically inspired free expression of their views through being too conscious of the institutional surroundings in which they find themselves.[13] I have also noticed that chat room messages inspired during a lecture can contain insights that may form the basis for the development of the student's own future theories. Since these insights arise while the lecture is in progress, they may, at the request of students, result in the lecture continuing on a different track or turning into a seminar. What is remarkable is the development of a rapport between the members of the group arising out of the instant messaging taking place concurrently with the lecture by students who are visible to each other in the form of their avatars; this would be impossible in the physical world since it would disrupt the lecture.

A university in the electronic world should be a place where academic life can take its course. Therefore, it is essential to develop land with buildings and other elements conducive to an academic atmosphere in Second Life. At various times art galleries have been set up, which I also make use of during physical lectures, going into the Academia and observing the exhibits with students (at present there is a gallery of twentieth-century art and another of photographs). There are also concert halls with performances of streamed and live music. This is made possible by advances in 3-D technology; what matters here is not communication or visual images but engagement in the electronic space.[14]

The electronic university changes the teacher-student relationship, starting with students creating avatars for themselves and adopting online names, which they use whenever they make contact. When students engage in academic activity, they are entitled to manage the buildings and grounds, but in so doing they assume responsibility for academic property. It is also important that when they visit Second Life, they are, to a certain extent, representing the Academia in particular and the academic world in general, which places certain obligations and responsibilities upon them.

The question of trust and responsibility is fundamental as it is concerned with the existence on the web of a university, a different kind of place and one that is respectful of the academic world. It is important for the university to observe the principle of openness (open lectures, events fostering an academic community, continuous access, and the opportunity for creativity), while at the same time maintaining its status. If it intends to exist as a university in the electronic world, where individuals create their own, often private worlds, realize their dreams, and occasionally experience that life to the fullest, then every effort must be made to create an appropriate space for them. It is important that the university be accepted in the electronic world, while at the same time becoming a point of reference and center for different kinds of activity brought to the electronic world by others. There is clearly a place for a university in the electronic world as there is for any kind of activity. When a university is transposed to the electronic world, certain features are bound to be different when compared with academia in the physical world. These changes result from the different ontological reality prevailing in the electronic world. For example, appearing in the form of an avatar affects interaction between individuals, while university buildings and lecture theaters need not resemble their physical counterparts at all, bringing an air of innovation to the conduct of lectures. This entirely new quality, based on electronic reality, arises instantaneously and in a manner requiring a particular response.

Since 2007, almost 200 students have officially completed courses and several dozen lectures have been given by invited academics. In addition, numerous artistic and popular educational events have taken place. I believe that these kinds of academic activities point towards the university of the future, which will be first and foremost a place rather than a mosaic of lectures.

In June 2012, two historic events in Polish e-learning took place at the Academia Electronica. June 6 saw the first public defense of a doctoral dissertation, titled *Computer Games in the Perspective of the Anthropology of Everyday Life*) by Radosław Bomba (RL)/Radel Bailey (SL), doctoral advisor Andrzej Radomski (RL)/An Redinamus (SL), Maria Curie-Skłodowska University, Lublin. On June 22, the first defense of a masters dissertation, titled *The Existence of Responsibility on the Web*, was made by Aleksandra Budzisz (RL)/Skrzydlatamara (SL), masters advisor Sidey Myoo, Jagiellonian University, Krakow. Both events were recorded and are available on the Academia website.

The Academia Electronica owes its existence to the engagement of those with no professional connection with the university but who give of their technical expertise to maintain its proper functioning, including the website.

Every Monday since 2007 (except during the summer vacation), I enter the electronic world for a few hours, halting my activities in the physical world. Activity in the electronic world can be directed toward any reality or person one wishes. These worlds are mutually exclusive with regard to their ontologies and how time is spent in them: the individual is of paramount importance; the worlds are secondary.

Notes

1. Au, *Making of Second Life*, vi.
2. Manovich, *Language of the New Media*, 290.
3. Kyrre and Olsen, *Becoming through Technology*, 40–61; Hörl, "Luhmann," 94–121.
4. Buechner, *Fictional Entities*.
5. Turkle, *Life on the Screen*, 73.
6. Krueger, *Artificial Reality II*.
7. Baron, *Always On*, 222.
8. Fleischmann and Strauss, "Interactivity as Media Reflection," 76.
9. Boellstorff, *Coming of Age*, 29.

10. Dyens, *Metal and Flesh*, 33.
11. Hershman-Leeson, *Raw Data Diet*, 249.
12. Academia Electronica. www.academia-electronica.net
13. Borgman, *Scholarship in the Digital Age*, 4, 65.
14. Riha, "Interactive 3-D Documentary," 100–102; Abelson, Leeden, and Lewis, *Blown to Bits*, 80–82.

Bibliography

Abelson, Hal, Ken Leeden, and Harry Lewis. *Blown to Bits. Your Life, Liberty, and Happiness After the Digital Explosion*, Addison-Wesley, Boston 2008.

Au, Wagner James. *The Making of Second Life: Notes from the New World*. HarperCollins e-books, 2008.

Baron, Naomi. *Always On: Language in an Online and Mobile World*. Oxford: Oxford University Press, 2008.

Boellstorff, Tom. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton, NJ: Princeton University Press, 2008.

Borgman, Christine. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: The MIT Press, 2007.

Buechner, Jeff. *Fictional Entities and Augmented Reality: A Metaphysical Impossibility Result*. In *Journal of Evolution and Technology* 22, no. 1 (2011), http://jetpress.org/v22/buechner.htm.

Dyens, Ollivier. *Metal and Flesh: The Evolution of Man: Technology Takes Over*. Cambridge, MA: The MIT Press, 2001.

Fleischmann, Monika, and Wolfgang Strauss. "Interactivity as Media Reflection between Art and Science." In *The Art and Science of Interface and Interaction Design, Vol. 1*, edited by Christa Sommerer, Jain Lakhmi, and Laurent Mignionneau. Berlin: Springer-Verlag, 2008.

Hershman-Leeson, Lynn. "The Raw Data Diet, All-Consuming Bodies, and the Shape of Things to Come." In *Database Aesthetics*, *Art in the Age of Information Overflow*, edited by V. Vesna. Electronic Mediations, Vol. 20. University of Minnesota Press, 2007.

Hörl, Erich. "Luhmann, the Non-trivial Machine and the Neocybernetic Regime of Truth." In *Theory Culture & Society* 29, no. 3 (May 2012).

Krueger, Myron. *Artificial Reality II*. Addison-Wesley Publishing Company Inc., 1991.

Kyrre, Jan, and Berg Olsen. "Becoming through Technology." In *New Waves in Philosophy of Technology*, edited by Jan Kyrre, Berg Olsen, Evan Selinger, and Søren Riis. New York: Palgrave Macmillan, 2009.

Manovich, Lev. *The Language of the New Media*. Cambridge, MA: The MIT Press, 2001.

Riha, Daniel. "Biography as an Interactive 3-D Documentary." In *Digital Memories Exploring Critical Issues*, edited by Daniel Riha and Anna Maj. Oxford: Inter-Disciplinary Press, 2009.

Turkle, Sherry. *Life on the Screen: Identity in the Age of the Internet*. New York: Simon & Schuster, 1997.

Paths to Defeasibility: Reply to Schauer on Hart

Ronald Loui
University of Illinois–Springfield

Frederick Schauer's attention has recently been drawn to defeasibility in a paper with a provocative title: "Is Defeasibility an Essential Property of Law?"[1] The crisis of confidence for Schauer appears to take hold about the time he reviews Nicola Lacey's biography of H. L. A. Hart.[2] Schauer actually finds room for defeasibility in a legal system, along the lines of judicial nullification of rule-derived legal guidance. He permits an ethical override of the logic and language of law, as a strongly desirable power granted the wise jurist in a system that is truly justice-seeking.

The most significant push for defeasibility has been felt in the community that has attempted to model legal reasoning with computer programs. The AI (artificial intelligence) and Law community, an international group of interdisciplinarians, visited the concept of defeasibility two decades ago. In fact, defeasibility has become so entrenched in AI and Law that the development of defeasible reasoning has advanced formally and mathematically within this milieux. Henry Prakken, for example, a lecturer in the Intelligent Systems Group of the computer science department at Utrecht University, and professor of law and IT at the Law Faculty of the University of Groningen, wrote his 1993 thesis at Free University Amsterdam, titled *Logical Tools for Modelling Legal Argument*. In 2002, he would be invited to write the review article "Logics for Defeasible Argumentation" with Gerard Vreeswijk for the *Handbook of Philosophical Logic*. Defeasible logic has also benefitted from the theses at Maastricht's law school by a mathematician, Bart Verheij, *Dialectical Argumentation with Argumentation Schemes: An Approach to Legal Logic,* and a computer scientist, Arno Lodder, *DiaLaw: On Legal Justification and Dialogical Models of Argumentation*. Verheij's advisor, Jaap Hage, a legal philosopher, added *Reasoning with Rules: An Essay on Legal Reasoning and Its Underlying Logic*. Those are just some of the Dutch researchers. Prominent proponents of defeasibility can be found in the AI and Law community from Italy, Argentina, Australia, the United Kingdom, Germany, France, Canada, Thailand, China, Japan, and the United States.

Apparently the desire to explain legal reasoning in enough detail that a computer system could be designed around the explanation has led many researchers to "dialogical defeasible argumentation," regardless of prior logical or legal tradition.

Yet, Schauer has apparently lost the will to defend the very defeasibility he found so interesting in his 1993 *Playing By the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Like H. L. A. Hart himself, who introduced defeasibility to the Western philosophical vernacular, then nearly disavowed defeasibility in the introduction to his collected works, there has been a noticeable retreat.[3]

Those of us tasked with designing actual systems of symbol manipulation that perform quasi-legal reasoning remain steadfast in our appraisal of defeasibility as a useful design paradigm. The purpose of this short note is to briefly review the main places in the analysis of legal reasoning where defeasibility finds its use.

Before I enumerate, it is worth remembering some history.

Defeasibility entered artificial intelligence and computer modeling in the storm that was "non-monotonic logic," an idea that occupied a Rockefeller-sized fraction of the AI field's intellectual investment at its peak. Rationalization of this situation came slowly, as epistemologically oriented philosophers such as John Pollock and Henry Kyburg began to weigh in. The philosophical tradition remains a moderating partner, while non-monotonic logicians, especially adherents to "default logic," continue with their creative flows. Pollock was influenced by Roderick Chisholm through John Ladd, but he always claimed he was trying to interpret Wittgenstein directly (although Waismann might be an equally good *locus focus* for Pollock's pre-formal work). Wittgenstein also inspired Jon Doyle, author of AI's truth-maintenance system, one of the major breaks from the attempt to do non-monotonic reasoning as a kind of modal belief logic.

In an era of renewed US interest in Constitution and secession, it is worth remembering that "indefeasible" was a popular high note of the classically trained rhetorician, especially when drawing a line in the sand: in the *1776 Virginia Declaration of Rights*, "community hath an indubitable, inalienable, and indefeasible right to reform, alter or abolish government . . ." (attributed to James Madison); and John Adams: "The people

have a right, an indisputable, unalienable, indefeasible, divine right to that most dreaded and envied kind of knowledge—I mean of the character and conduct of their rulers." Also, Lord Aberdeen: "indefeasible right inherent in the British Crown" and Gouverneur Morris: "the Basis of our own Constitution is the indefeasible Right of the People." Scholarship about Abraham Lincoln often cites these passages in the justification of secession.

If "indefeasible" was necessary as adjectival qualifier for the country's first one hundred years, it is because the normal situation was the defeasibility of rules. That is, defeasibility was the default. In recent years, after the wide influence of deductive logical thought and perhaps concomitant extremist political orientation, the situation has reversed. A rule must be adjectivally qualified as "defeasible" to reassure the rule-governed that there is recognition of potential exception. Defeasibility is now the exception.

So how does the computer programmer get drawn to defeasibility?

Convention: rule qualification, rule emendation, and rule priority

The rule "if p, then q" admits an exception, "if r and p, then not q," because sometimes, rules simply require qualification. It is a fool's errand to think that "if not-r and p, then q" exactly mirrors the situation. Such a false transformation has been discussed endlessly in multiple literatures (even Hart notes the non-equivalence in a footnote). Those who construct conclusions through procedures, or who use multi-valued logical systems, or context, to model a belief ascription or prescription, have ample room to make a distinction.

It is simply an empirical fact that rules are asserted that are then emended with qualifiers and undercutters. Those who want to constantly rewrite into a two-valued system not only lose the force of the rule when r's assertibility, knowability, or believability is unknown; they also lose the naturalness of taking the declaration, "if p, then q," as a logical conditional, "if p, then q," at face value, because they must re-represent the former as a rule with complex (and always unfinished) antecedent. The resulting conditional looks nothing like the plain language declaration. In a choice between two logical conventions, the deductive approach is less convenient than the defeasible.

As a trump card, rules in judicial systems are produced within jurisdictions. *Lex superior* remains part of ex-Latinate legal language for a reason. When a superior court applies its rules to countermand the rule-governed conclusion of a lower court, or when federal law simply nullifies that of a rogue Southern state, we have defeasibility. Rule priority can certainly be modeled with lots of different order-producing structures in mathematics, but the lower-level derivations must be given some kind of description pertaining to their potential to be vacated: some prefer *provisional* or *prima facie*, though both of these descriptions are too weak to capture the fairly calculated and fully invested, authoritative conclusions of a lower court.

Another quick hit for convention: linguistically, rule-givers often give rules with their own priority meta-rules for resolving conflict; they create their own rule-priority strata even without jurisdictional complication.

Assertion and argumentation: a third truth value

The idea of a third value is generally attractive to those who want more flexibility than a modal logic of belief. In the modeling of the give and take of dialectical or dialogical argument, it is *Peisione, Agalope, and Thelxiepeia*. If an argument applies a rule, non-demonstratively, the rule's implication must be described. What status should it be given? Again, is it "provisional"? "Fallible and corrigible"? "Asserted and not yet rebutted?" "The output of a non-demonstrative line of argument based on subsumption under non-demonstrative rules, which may be demoted through further argument"? The latter is exactly what we mean by "defeasible."

Deductive approaches to dialectical argument seek to recognize a set of assertions that are "advanced but not rebutted" along with the material conditionals from which they might be derived. Suppose "p" is in. Suppose "if p then q" is in. If you dispute "q," then which of the two earlier sentences do you reject? This is the medieval *obligation game* of course, and dialogical models of argument, "dialogue logics," continue to be developed along these lines. But the moment we reason about "p," or even "if p then q," at the meta-level, there are rules that decide whether or not the sentence has the "in" status. These rules are what? Defeasible? Or is there always a logically consistent set of meta-assertions that are "meta-in" with meta-disputation over meta-rules, with a meta-meta-level, a meta-meta-meta-level, and a regress to infinity?

As a practical matter, we prefer our infinities to be non-terminating processes, not indefinite representations. That's because we would actually like to start our computations, even if they run forever, rather than spend forever trying to specify our computations.

Language: the lazy learning of open texture, incommensurability of language, elision of detail, and legislative compromise

The two remaining main roads to defeasibility are less obvious, but carry more traffic. The first of the two has to do with the semantics of words used in rules and natural language arguments. Hart famously talked about the open texture of terms in law. Joel Feinberg called it the accordion. In computational learning theory, they call it "lazy learning." The idea is this: the cutting planes in high-dimensional space that distinguish positive from negative examples are not yet defined, or are only grossly stated, on first linguistic encounter. The speaker and hearer understand that future distinctions will be made as hard cases arise, by subjecting the linguistic community to procedures (possibly non-deterministic procedures) that are well determined.

As I once asked Supreme Court Justice Scalia, "Do you really think there must be agreement on the meaning of the Constitution, when it would suffice that there is agreement on how to resolve disagreements on the meaning of the Constitution?" (He responded, "yes," but that answer paradoxically undercuts the authority of his own answer.)

Language changes, and so too does the rationale of rules, their intention, and their compact or efficient expression. Frequently, rules are expressed in a way that elides a more intricate argumentative basis.[4] H. L. A. Hart's famous example of "no vehicles in the park" could not have anticipated unmanned combat drones flying overhead during military parade, in the airspace over a park that has been purchased by a citizen preservation council and leased back to the government under a legally novel contract.

Rules are especially subject to deprecation or derogation when their legislative origins are known to have been conflicted. In such cases, the original openness of terms may be significant, the hard work being left to the courts to blaze a trail of precedent (which trail is itself subject to future legislative revision or annulment).

When rules are open textured, newly decided cases pin down the semantics in regions where meaning may have been undetermined, unclear, or muddled. But new cases may also slice away at regions where meaning appeared fixed by default,

thus defeating prior inferences and providing a stronger kind of defeasance.

Some may think that lazy learning is the result of lazy speaking, that precision of meaning is possible for those whose discipline and diligence provide for superior diction. This is a chimaera of idealist thought. Even mathematicians resort to defeasible specification as a matter of necessity in their language: The Dirac delta function is "a generalized function on the real number line that is zero everywhere except at zero, with an integral of one over the entire real line." This is not disputed. What is really outrageous in the practice of actual mathematics is the production of "proofs" in mathematical papers that are taken to be correct until "shown" otherwise.

Reason: analogical reasoning from precedent

The fourth path to defeasibility is perhaps the most widely accepted because there are no good competing formal models. Analogical reasoning has always been known to be non-deductive, non-demonstrative, and ampliative, requiring something more than a Fregean-Russellian logic for its description. Analogical reasoning in law, especially common law, is crucial for those who would model logical thought as a processing of symbols. Here is an area where AI and law had recently delivered substantially.

There are other non-demonstrative forms of reasoning, besides analogical reasoning, that do not appear to benefit from representations using defeasible rules or a defeasible truth-status. For example, Kyburg formalizes scientific theory induction as a shift of the meaning postulates governing theoretical terms. His shift is revisable but not defeasible (each inductive revision is the best that is possible at the time, not subject to revision through further adversarial process, further derivation, or further computation of any kind).

Our model of analogical reasoning from the prior case starts by representing the arguments that were produced in that case. One can see analogies to the case, with its prior arguments, together with the prior judgments as to which arguments were persuasive in the prior case, producing "provisional" conclusions in a new fact situation. Each relevant case can take the facts as input and generate some conclusion as output; each case is a fact-conclusion-generator. One can further derive defeasible rules from such a case, and represent them as first class objects. Obviously, by giving these fact-conclusion-generators a name, by calling each a defeasible rule, there can be meta-reasoning about them, which provides more nuance. Hence, we refer to the rules of the case in *Ladue v. Gilleo* and *Ward v. Rock Against Racism*. We do not just carry around ten million black boxes that take input fact situations and produce "provisional" conclusions, then deal with the outputs as if they were all on equal footing, with no internal structure.

The resulting logic of analogy, using defeasible rules of precedent cases, dwarfs the earlier account of analogy in law by Joseph Raz as much as modern risk portfolio analysts on Wall Street make Pascal look like a newbie.

These are the four main reasons why I find defeasible conditionals and a defeasible propositional attitude helpful in the production of logics of legal reasoning that have functionality, systemic integrity, and conventional naturalness. Other designers may have different subjective appraisal. They may be loyal to former logical training. They may simply have different interests. The proof will be in the depth of future developments following different symbol system paradigms.

AI and Law has its own skeptics regarding defeasibility. The founder of the field, L. Thorne McCarty, wrote a dissent, "Some Arguments about Legal Arguments," pointing out that most legislators try to enumerate exceptions as antecedent conditions when they pen their social rules.[5] Later, McCarthy would allow that defeasibility and deontic logic make a nice match. There remains a strong school of thought that holds belief revision and paraconsistency, not defeasibility, to be the keys to formalizing logics of legal reasoning. Even Fred Schauer appears to have found himself among skeptics, among the deontic logic students of Carlos Alchourrón. Alchourrón famously struggled to resist defeasibility on his deathbed.[6] As someone who has attempted to model intricate and subtle patterns of reasoning as formal symbol systems, I wish them well, while seriously doubting that they can achieve the elegance and compactness that we have found with defeasible logics in our AI and Law models. It may sound pithy, but unlike Schauer and Hart, my support for defeasibility remains indisputable, unalienable, and indefeasible.

Notes

1. Schauer, *The Logic of Legal Requirements: Essays on Defeasibility*, ed. Jordi Ferrer Beltran and Giovanni Battista Ratti (Oxford University Press, 2012).
2. Schauer on Lacey on Hart, "(Re)Taking Hart," *Harv. L. Rev.* 119 (2006): 852, although see also "On The supposed Defeasibility of Legal Rules," *Current Legal Problems* 51, no. 1 (1998): 223–40.
3. Schauer was actually my keynote speaker at the AI and Law conference in 2001 in St. Louis, at the time the director of the John F. Kennedy School of Government Institute of Politics and senior constitutional law professor at Harvard, a speaker in the Abe Lincoln *Hurd et al. v. Railroad Bridge Co.* historic court room, my replacement for a little known Illinois Senator named Obama.
4. See R. P. Loui and Jeff Norman, "Rationales and Argument Moves," *Artificial Intelligence and Law* 3, no. 3 (1995): 159–89.
5. McCarty, "Some Arguments about Legal Arguments," *ICAIL* (1997): 215–24.
6. R. Loui, "Alchourrón and von Wright on Conflict among Norms," in *Defeasible Deontic Logic*, ed. D. Nute (Dordrecht: Springer, 1997).

Bibliography

Adams, John. *Thoughts on Government*, 1776. Hayes Barton Press, 2008.

Anastaplo, George. *Abraham Lincoln: A Constitutional Biography*. Lanham: Rowman & Littlefield, 2001.

Doyle, Jon. "A Truth Maintenance System." *Artificial Intelligence* 12, no. 3 (1979): 251–72.

Feinberg, Joel. "Action and Responsibility." In *Philosophy in America*, edited by Max Black. London: Allen & Unwin, 1965.

Hart, H. L. A. *The Concept of Law*. Oxford: Clarendon Press, 1961.

Krabbe, E. C. W. "Dialogue Logic." *Handbook of the History of Logic vol. 7*. Elsevier, 2006.

Kyburg, Henry. *Science and Reason*. Oxford: Oxford University Press, 1990.

Madison, James. *The Writings of James Madison: 1787–1790*. New York and London: G. P. Putnam's Sons, 1904.

Raz, Joseph. *Practical Reason and Norms*. Princeton: Princeton University Press, 1975.

Computational Philosophy and the Examined Text: A Tale of Two Encyclopedias

Colin Allen
Indiana University, Bloomington

Jaimie Murdock
Indiana University

Cameron Buckner
University of Houston

Robert Rose

[This piece for the APA Newsletter on Computing and Philosophy is an abridged version of the keynote address delivered by the first author at the joint AISB/IACAP meeting in Birmingham, England, in July 2012. The meeting honored the centenary of Alan Turing's birth.]

What might philosophers do with millions of words? Read them, of course. But the days of being able to cover everything are long gone. Even online encyclopedias that provide considerable compression of the primary and secondary literatures are at the limits of what can be reasonably assimilated. The Stanford Encyclopedia of Philosophy (SEP) continues to grow past 13 million words, while the Internet Encyclopedia of Philosophy (IEP) comprises another 4 million. The PhilPapers database contains half a million records of article titles, along with abstracts for about half of them, which adds up to a whopping 21 million words even before one starts to look at the articles and books themselves. At roughly a page a minute, it would take almost 2,500 hours, or a year of more than full-time reading to read through these three tertiary sources: a Borges-ian task, from which one could expect to retain only isolated memories. Those who manage these growing collections also face a never ending task of tracking the dozens of new articles appearing daily, staying on top of new thematic developments, and identifying the connections to existing materials.

Clearly, philosophers need good information management systems, and at the Indiana Philosophy Ontology (InPhO) project we have been working to provide infrastructure for a variety of philosophical applications, making all of our data available through a human-friendly interface (https://inpho.cogs.indiana.edu/) and a machine-friendly application programmer interface (https://inpho.cogs.indiana.edu/api). But that's the relatively boring stuff to most philosophers. More exciting is what philosophers can learn about their discipline and its texts from these kinds of tools. Our goal with the InPhO project is to provide a platform for systematically investigating various philosophical texts and interrogating the different computational models that can be applied to them. Here we describe some of our preliminary investigations and discuss the next steps to take.

Starting with the three sets of text already mentioned—the SEP, the IEP, and PhilPapers—we are exploring their similarities and differences. For instance, as a practical issue, one might want to know whether the IEP and the SEP have equivalent coverage of different philosophers and their ideas. If one knows the relevant parts of both encyclopedias already, one may suspect that their coverage is not the same, but can we measure and represent the ways in which they diverge? Computer scientists and computational linguists have developed a number of statistical methods for extracting networks of related terms from large bodies of text, but in these disciplines the goal is usually to show that one's favorite algorithm outperforms its rivals. In other words, models are pitted against each other, and then their performance is measured against some "gold standard." However, in our case we have no gold standard. Nevertheless, we believe there is much to learn about philosophical texts by mapping both where the models agree and where they disagree.

A full analysis of the different models we have implemented is beyond the scope of this short piece. Instead, we describe some aspects of the behavior of one algorithm applied to the resources mentioned above. The modeling approach is the "bound encoding of the aggregate language environment" (or "BEAGLE") model of Jones & Mewhort (2007). Our selection of this model is intended to be illustrative only—we are actively investigating other models—but we like the BEAGLE model because it combines word context and word order information into a holographic representation of the corpus, and it has proven successful as a cognitive model of people's intuitive judgments about semantic similarity among words. The term similarities within and between corpora can be used to help solve some of the information management issues mentioned above, as well as helping us understand the different themes and emphases among the various corpora.

The BEAGLE system is "trained" on a corpus by a process that iteratively builds up a vector of n bits (where n is a parameter) representing every term in the corpus. The vectors representing each term are initially randomized, but the algorithm produces vectors that cluster in n-dimensional space in ways that reflect the underlying semantics of the terms. We trained separate instances of BEAGLE on the SEP corpus and the IEP corpus, and, for reasons to be explained below, we trained a third instance of the model on the combined SEP and IEP corpus. Similarity between vectors in a multidimensional space can be (crudely) represented by the cosine value between them, and the resulting set of cosines provides constitutes m x m matrix, where m is the number of unique terms in the corpus (approximately 200,000 for the SEP and approximately 120,000 for the IEP) that is too large to render visually or summarize succinctly. Nevertheless, but probing this matrix systematically, it is possible to learn something about the two encyclopedias.

Three examples

Suppose we are interested in knowing how the SEP and the IEP compare in their treatment of Kant, or Aristotle, or metaphysics, or ethics. A simple approach is to use the cosine values to extract and compare the most similar terms between the two. Thus, for example, the BEAGLE model applied to the SEP identifies the most similar terms to "Aristotle" as "plato," "metaphysics," "boethius," "alexander," "avicenna," "physics," "works," "porphyry," "taking," "topics," "augustine," and "descartes," while for the IEP the "aristotle" list consists of "plato," "ethics," "kant," "unlike," "metaphysics," "spinoza," "leibniz," "descartes," "despite," "contemporary," "aquinas," and "taking." Clearly there is some overlap here, but also differences (as well as some apparent noise from terms such as "taking"). How can we understand these?

One simple approach is illustrated in Figure 1, which shows the 10 most similar terms the two encyclopedias as modeled by BEAGLE for four key words: "aristotle," "metaphysics," "epistemology," and "ethics." The links are colored in red if the link comes from the SEP model only, blue if from the IEP only, and purple if from both, and the wider links correspond to higher cosine values (0.7 or greater) with the purple links representing the averaged cosine value in the two trained models. It is plain to see that there is more overlap around "metaphysics," and the cosine values are typically higher, than around the other terms, although one must be cautious about drawing strong conclusions from such a limited analysis.

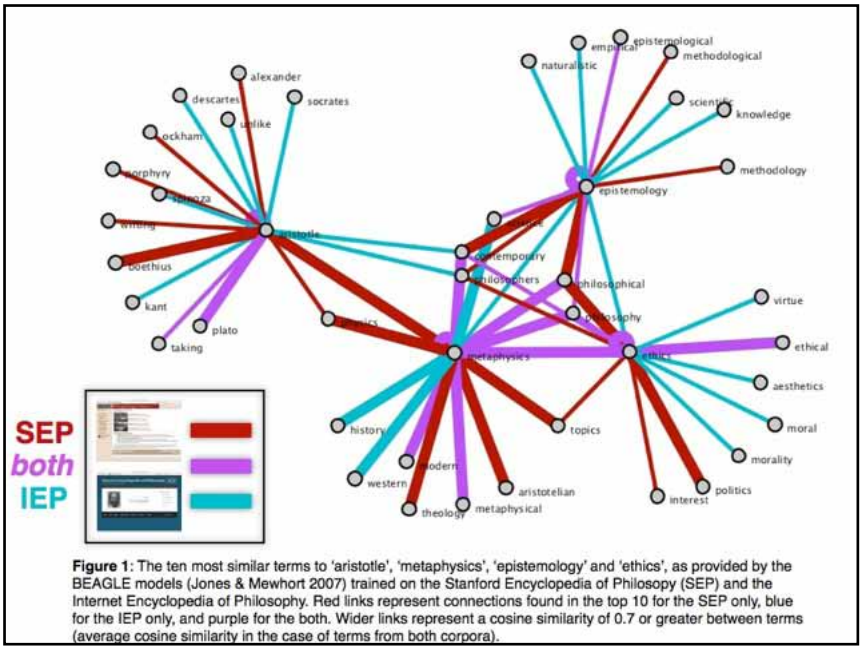


Figure 1: The ten most similar terms to 'aristotle', 'metaphysics', 'epistemology' and 'ethics', as provided by the BEAGLE models (Jones & Mewhort 2007) trained on the Stanford Encyclopedia of Philosophy (SEP) and the Internet Encyclopedia of Philosophy. Red links represent connections found in the top 10 for the SEP only, blue for the IEP only, and purple for the both. Wider links represent a cosine similarity of 0.7 or greater between terms (average cosine similarity in the case of terms from both corpora).

Similarly interesting are the different neighborhoods around "kant" with the BEAGLE model of the SEP yielding the philosophers names "hume," "leibniz," "herder," "cohen," "locke," and "wolfe" among the top twenty terms, while the IEP-trained instance of BEAGLE has "hegel," "aristotle," "leibniz," "descartes," "Schopenhauer," "maimon," "fichte," "hume," and "Spinoza" among its top twenty.

But if these differences show up when the top twenty terms are considered, what if we look at fewer, or more? One way to measure the degree of agreement or disagreement between lists is to use the Spearman rank coefficient derived from the positions of items found on both lists. Two identical lists will have a Spearman coefficient of 1, a list and its reverse have a coefficient of -1, and if one list has random order with respect to the other the correlation coefficient will be 0. Over the twenty most similar terms to "kant," the SEP and IEP have a Spearman coefficient of 0.764 and with just ten terms it is even higher at 0.806, whereas at fifty terms, the coefficient has dropped to 0.326. However, this is not a simple decreasing function with correlation dropping off the more terms are considered, as the graph in Figure 2 shows (the correlation coefficients for "metaphysics" are plotted for comparison). In fact, there's a double dip and a steady rise as more and more terms are considered, with the correlation between the two encyclopedias around the term "kant" having reached again 0.737 when the neighborhood is expanded to 7,000 terms (the correlation coefficient for "metaphysics" is even higher at 0.763 for 7,000 terms).

We leave a more complete analysis of this for another time, but a couple of general comments are in order. First, this preliminary investigation suggests that the double-dip phenomenon with a long rising tail may be a common profile when rank coefficients are plotted using the BEAGLE similarity data from the two encyclopedias. However, the depth and precise location of these dips vary for different terms. Our interpretation of this phenomenon, if it is indeed robust, suggests that reference works will agree on a few most highly relevant items, disagree on matters of emphasis for many of the moderately related items, but will both mention nearly everything the other does at the periphery. This circumscribes the degree of freedom that reference works have to influence professions through differences of emphasis. From a certain vantage point, the IEP coverage of Kant may seem more skewed towards his German idealist successors than the SEP coverage, and the SEP coverage more skewed towards Kant's place in the rationalist-empiricist debate. However, by taking a narrower (top ten most related terms) or broader (top 500), those differences may not be so evident . . . or important.

To reduce everything to a single number, the rank coefficient, representing the correlation between two ordered lists is, of course, to throw away a lot of the structure that is present in each of the cosine matrices (one from each instance of the trained model) from which the lists (e.g., the closest associates of "kant" in each trained model) were extracted. Currently at InPhO we are also trying to develop ways of visualizing this structure. In particular, as well as the first order relationships between a given head term (e.g., "kant") and other terms in the corpus, the BEAGLE model (like any other vector model) provides cosine values between each pair of terms in the list. Such a matrix of cosine values can be represented as a network that can be laid out by a clustering algorithm that attempts to preserve as much of the multidimensional distance information in a two-dimensional format as possible—the so-called multi-dimensional scaling (MDS) algorithm. Multi-dimensional scaling enables one to visualize high-dimensional data in a two dimensional "map," where distance in the map represents dissimilarity of two datapoints in the original high-dimensional space. Using MDS it is possible to generate two separate maps of term relations from each of the models of the SEP and the IEP, but it is virtually impossible for the human visual system to extract anything interesting from such side-by-side displays.

Instead, we borrow a trick from Kievit-Kylar & Jones (2012) where we first layout out a network in a semantically significant way that is relatively neutral between the two sources of text. This is then used as a background against which to highlight the SEP and IEP networks. The "neutral" representation is

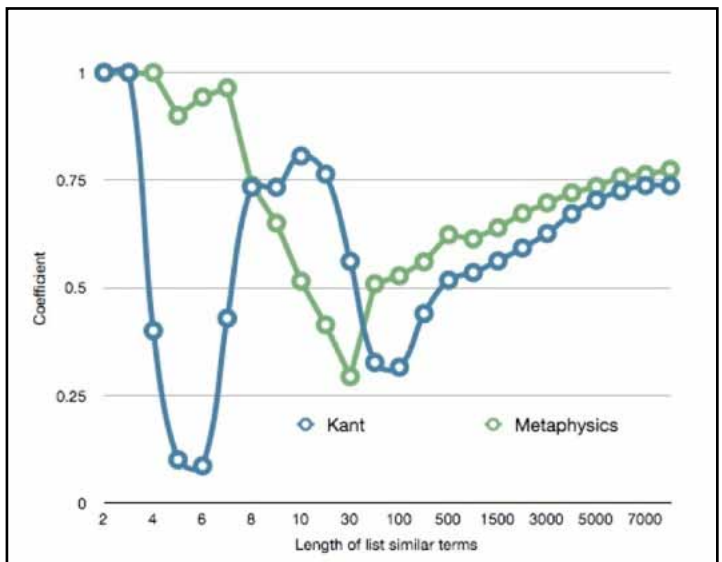


Figure 2: Spearman rank coeffiecients (y axis) computed from lists of terms most similar to 'kant' and 'metaphysics' from BEAGLE models (Jones & Mewhort 2007) trained on the SEP and the IEP, plotted against the number of most similar terms (x axis) selected for the comparison. (Note the nonlinear scale on the x axis.)

obtained by training a third instance of the BEAGLE model on both encyclopedias simultaneously. For this demonstration, we picked "turing" as our head term. We formed the union of the top 100 most related terms from each of the independently trained models, which yielded about 140 terms, and then we manually discarded a few terms that seemed to be uninteresting verb forms. We then used MDS visualization of these 128 terms as represented in the joint SEP-IEP BEAGLE model to provide the background against which the single-encyclopedia models could be compared. By highlighting those terms judged most similar to each other in the SEP-only model in red, and those judged most similar to each other in the IEP-only model in blue, and filtering out relatively weak connections (low cosine values) we get the result shown in Figure 3. Here, some differences really stand out with, for example, the SEP showing stronger linkages among terms associated with formal computational theory, while the IEP's strongest links are among terms relating more general discussions of human intelligence and rationality, although there is also a common core shared by the two encyclopedias.

Again, we caution that a full analysis of these results must be more systematic. But for now we are encouraged and excited by the potential of our methods to yield insight into the way that philosophers write about their subject, and the value of different representations of ideas, thinkers, and their relations.

Future work

We have also applied these modeling techniques to the PhilPapers database *and recently acquired access to approximately 3 million volumes of the Google Books/Hathi Trust collection*. Issues of scale arise in the latter case, and progress will be slow. However, we know already that by bringing PhilPapers into the mix, we can help clarify questions about the differences between *how* the SEP and the IEP represent philosophy.

Clearly, there are more experiments possible with our data than any of us can conduct. Therefore at the InPhO we are committed *to open access to our data and open source to provide both the raw materials and the tools* that will allow others to conduct similar experiments. To that end we have recently opened out InPhOSemantics DataBlog at http://inpho.cogs.indiana.edu/datablog/ and are inviting philosophers everywhere to experiment and to share their analyses and visualizations of these data.

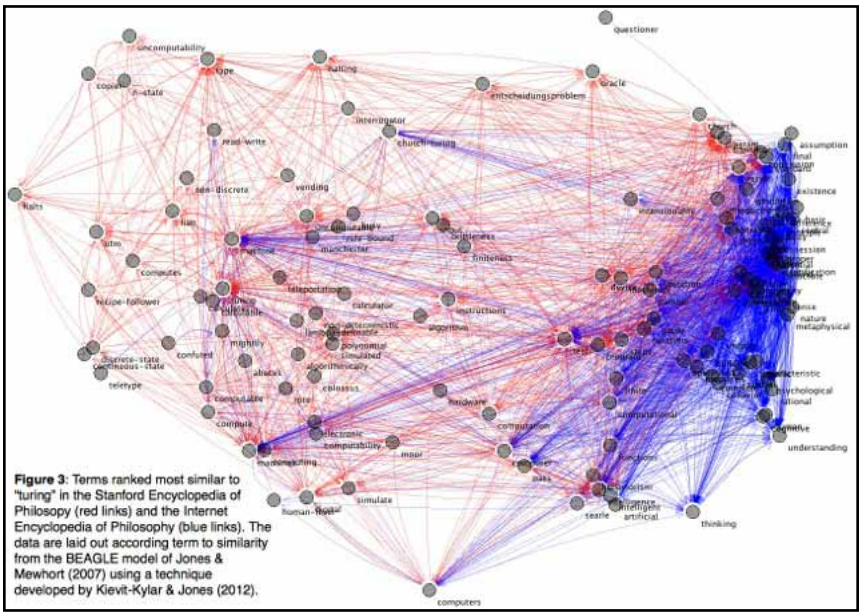


Figure 3: Terms ranked most similar to "turing" in the Stanford Encyclopedia of Philosophy (red links) and the Internet Encyclopedia of Philosophy (blue links). The data are laid out according term to similarity from the BEAGLE model of Jones & Mewhort (2007) using a technique developed by Kievit-Kylar & Jones (2012).

Acknowledgments

We would like to thank all associated with the InPhO project for their help, especially Jun Otsuka for providing comments on an earlier draft. We are grateful to David Bourget for providing access to the PhilPapers (http://philpapers.org/) database, and to the editors of the Stanford Encyclopedia of Philosophy (http://plato.stanford.edu/) and the Internet Encyclopedia of Philosophy (http://www.iep.utm.edu/) for their commitment to open access. Our research would not be possible without generous support from the National Endowment of Humanities, grant numbers PW-50401-09 and HJ-50092-12. We would also like to thank the Indiana University Cognitive Science Department for material and personnel support.

References

Buckner C., Niepert M., Allen C. "From Encyclopedia to Ontology: Toward Dynamic Representation of the Discipline of Philosophy." *Synthese* 182 (2011): 205–33. http://dx.doi.org/10.1007/s11229-009-9659-9.

Jones, M. N., and D. J. K. Mewhort. "Representing Word Meaning and Order Information in a Composite Holographic Lexicon." *Psychological Review* 114 (2007): 1–37. http://dx.doi.org/10.1037/0033-295X.114.1.1.

Kievit-Kylar, B., and M. N. Jones. "Visualizing Multiple Word Similarity Measures." *Behavioral Research* 44 (2012): 656–74. http://dx.doi.org/10.3758/s13428-012-0236-7.

What We Can Learn from the Failure of the Singularity

Federico Gobbo
University of L'Aquila, Italy

The history and philosophy of AI (artificial intelligence) has kept inspiration from science fiction since the beginning. Sometimes epistemological and ethical issues arise in novels before they appear in scholarly essays—think, for example, about Isaac Asimov's works and roboethics.

Bruce Sterling is the founder—together with William Gibson—of the cyberpunk avangarde movement. In the late 1970s, a group of artists (mainly writers, but also mathematicians and computer scientists) envisaged the importance of connected networks in our daily lives. In 1984 he published *Neuromancer*, where a "console cowboy" (i.e., a computer hacker) acted mainly in the cyberspace, that is the Net, not far from what we know today. That year was crucial in the history of computing. For instance, think about the Orwell's dystopia *1984*, the commercial that launched the first Apple computer, the collapse of the home computing market and the start of the GNU project, the nucleus of Richard Stallman's free software movement.

Recently, recalling that 2013 is the thirtieth anniversary of the essay by Vinge where the concept of Singularity was proposed, Sterling argued, in the columns of the *Edge.org*:

> This aging sci-fi notion has lost its conceptual teeth [...] It's just not happening. All the symptoms are absent. Computer hardware is not accelerating on any exponential runway beyond all hope of control. We're no closer to "self-aware" machines than we were in the remote 1960s. Modern wireless devices in a modern Cloud are an entirely different cyber-paradigm than imaginary 1990s "minds on nonbiological substrates" that might allegedly have the "computational power of a human brain." A Singularity has no business model, no major power

> group in our society is interested in provoking one, nobody who matters sees any reason to create one, there's no there there. [. . .] We're getting what Vinge predicted would happen without a Singularity, which is "a glut of technical riches never properly absorbed." There's all kinds of mayhem in that junkyard, but the AI Rapture isn't lurking in there.

In spite of Sterling's good sense, the AI Rapture has still adepts nowadays. Chalmers (2010) has even proposed an analysis of its consequences, as if the Singularity were just around the corner. For the author, AI means that machines are at least equivalent to our knowledge, then he extends the concept introducing "AI+" (i.e., where machines will be more intelligent than the most intelligent human beings), while "AI++" is the ultimate amplification, where the intelligence of the machines will be so great that we will feel like mice in the presence of them—here the Singularity begins. The line of reasoning is summarized as follows:

(i) If there is AI, AI will be produced by an extendible method.

(ii) If AI is produced by an extendible method, we will have the capacity to extend the method (soon after).

(iii) Extending the method that produces an AI will yield an AI+.

(iv) Absent defeaters, if there is AI, there will (soon after) be AI+.

Chalmers's arguments contain a mistake and a fallacy. The mistake is in the use of AI itself: AI is not what Chalmers claims; rather, it is the complex of efforts *towards* the goal (i.e., to build at least a single machine as intelligent as us). It suffices to open the door of a university department where AI is the main topic somewhere in the world. As advocated by Müller (2010):

> Overall, the theory and philosophy of AI has set itself free from the single focus on the criticism of computational symbol manipulation; it has moved towards a new Cognitive Science and, in some quarters, a less intimate link with Cognitive Science overall. These developments support a more constructive cooperation with those who do "the real work"—but also face the real basic problems.

But let us accept the immediate objection: for the sake of the argument, it is not relevant when it will happen, or even *if* it would happen. Following the objection, we should then accept Chalmers' definition of AI. The corollary is that now we are in the stage of AI-, that is our "intelligent" machines are mice compared to us. In a sense, we experience an Anti-Singularity. This is for the mistake. Now, we turn to the fallacy.

Dreyfus (2012) adverted us that the extension method—used by Chalmers, among others—that is, we build AI, then AI+, and finally AI++ relies on the first step fallacy: we have no guarantee that the success in the first step of a construction will lead us to the last step. In our world of AI-, we are still learning how to make the first step.

Overall, it seems to me that we can invert the Virtuality Fallacy (recently proposed by Moor, in Tavani 2010, ch. 3) to give an account of the paralogisms followed by the AI Rapture, and in particular the Singularity enthusiasts. This is my proposal for the Inverted Virtuality Fallacy:

(i) X exists in virtual.

(ii) Cyberspace is virtual.

(iii) X (or the effect of X) is real.

I think that philosophers should work side by side with people who do "the real work," as Müller in the quotation above said, because we all have not only duties but also responsibilities of helping humanity to understand (epistemological first step) and improve (ethical second step) the world we are living in now and tomorrow, instead of speculating about possible worlds that most probably will never occur. If you are really interested in it, then write a good science fiction story.

Bibliography

Chalmers, David. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (2010): 7–65.

Dreyfus, Hubert L. "A History of First Step Fallacies." *Minds & Machines* 22 (2012): 87–99.

Müller, Vincent. "Philosophy and Theory of Artificial Intelligence." *The Reasoner* 5, no. 11 (October 3-4, 2011): 192

Sterling, Bruce (2012). "'The Singularity': There's No There There." *Edge.org*.

Sterling, Bruce. *Neuromancer*. Ace, 1984.

Tavani, Herman T. *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*, 3rd ed. John Wiley & Sons, 2007.

Vinge, Vernor. "The Singularity." Originally published in 1993 as an academic paper. Reprinted in *Whole Earth Review* (Spring 2003).