Philosophy and Phenomenological Research
Vol. LIX, No. 1, March 1999

The Zombie Attack on the Computational Conception of Mind*

SELMER BRINGSJORD
Rensselaer Polytechnic Institute

Is it true that if zombies—creatures who are behaviorally indistinguishable from us, but no more conscious than a rock—are logically possible, the computational conception of mind is false? *Are* zombies logically possible? Are they *physically* possible? This paper is a careful, sustained argument for affirmative answers to these three questions.

1 Introduction

Many proponents of computationalism,[1] the view that cognition is computation, are busy trying to practice what they preach: they are trying to build artificial persons. Two such proponents are the philosophers John Pollock and Daniel Dennett. In his last two books, *How to Build a Person* [45] and *Cognitive Carpentry: A Blueprint for How to Build a Person* [44], Pollock argues that in the future his OSCAR system will be a full-fledged person. For Dennett, the person-to-be is the robot COG, or a descendant thereof, a being taking shape with Dennett's help at MIT.[2] I have advanced a number of arguments designed to establish that the "person building project" will inevitably fail, but that it *will* manage to produce artifacts capable of excelling in the famous Turing Test, and in its more stringent relatives.[3]

* For trenchant comments on ancestors of this paper I'm indebted to three anonymous referees (whose insights were especially helpful), John Searle (whose seminal discussion of zombies in his *The Rediscovery of the Mind* provides the first round of my ammunition), Daniel Dennett, Stevan Harnad, David Rosenthal, Robert Van Gulick (who offered particularly insightful comments on the remote ancestor presented at the 1994 Eastern APA Meeting), Peter Smith, Kieron O'Hara, Michael Zenzen, Jim Fahey, Marvin Minsky, Larry Hauser and Pat Hayes. David Chalmers provided helpful analysis of a previous draft, and I profited from reading his *The Conscious Mind*, wherein zombies are taken to be logically possible. Conversations with Ned Block and Bill Rapaport also proved to be valuable.

1 Sometimes also called 'Strong Artificial Intelligence' (Russell and Norvig: [51]), or 'GOFAI' (Haugeland: [37]), or 'the computational conception of mind' (Glymour: [34]), etc.

2 Dennett shares his vision in [24].

3 The first wave of my arguments are found in the monograph [14]. A refined and sustained argument for the view that Pollock, Dennett, and like-minded people will manage to produce non-persons capable of passing the Turing Test and its relatives can

What sort of artifacts will these creatures be? I offer an unflattering one-word response: Pollock, Dennett, and like-minded researchers are busy building ... *zombies*.[4]

Is it really possible that what Pollock and Dennett and other computationalists are building is a creature whose overt behavior is as sophisticated as ours, but whose inner life is as empty as a rock's? I believe so. I also believe—for reasons to be specified below—that the mere *possibility* of zombies is enough to explode the computational conception of mind.

A recent clash between Daniel Dennett and John Searle over zombies provides a tailor-made springboard to a sustained defense of the zombie attack against computationalism. Dennett, more than any other thinker, says that no philosopher of mind has anything to fear from zombies; in fact, he thinks that those philosophers who seriously ponder zombies (and Blockheads, Twin Earthlings, and Swampmen) have "lost their grip on reality" [22]. Searle, on the other hand, believes that zombies threaten at least behavioral conceptions of mentality. In this paper I try to show that Searle is right, and that he has laid the foundation for a new, rigorous attack on computationalism—the zombie attack. If this attack is sound, it will follow not only that aspiring person builders will fail, but that in failing they may indeed give us zombies.[5]

This paper is structured as follows. In section 1 I focus the Dennett-Searle clash, and then argue that Searle seems to be the immediate victor. In section 2 I adapt the results of section 1 so as to produce a disproof of computationalism. In section 3 I defend this disproof by destroying rebuttals from, and on behalf of, Dennett, including one from his *Consciousness Explained* which seeks to exploit David Rosenthal's "higher order theory" of consciousness. In section 4 I consider and reject two final rejoinders, one of which presses the question, "Well then, why aren't *we* zombies?" I end in section 5 with a brief summary.

1 Dennett's Dilemma

Dennett is the arch-defender of the computational conception of mind that underlies the "person building project"; Searle, on the other hand, is the arch-attacker—and both relish their roles: Dennett, in a rather harsh review [25] of

---

be found in [13]. New formal arguments against the person building project can be found in my [8] and [9]. Alan Turing presented his famous test in [59]. Stevan Harnad was the first to suggest more stringent systematic variants on the original Turing Test; see his [36].

4 I refer to *philosophers'* zombies, not those creatures who shuffle about half-dead in the movies. Actually, the zombies of cinematic fame apparently have real-life correlates created with a mixture of drugs and pre-death burial: see [20], [19].

5 When I refer to 'person builders' I refer to those who intend to replicate human persons in a *computational system*. Presumably there are more "biological" ways of striving to build persons—ways involving, e.g., cloning.

Searle's recent *The Rediscovery of the Mind* (= RM [52]), affirms that, from the perspective of Searle and like-minded anti-computationalist thinkers, he is the "enemy," and the "target representative of [cognitive] orthodoxy." Searle, as is well known (from his Chinese Room Argument [54]), and well-revealed repeatedly in RM, regards computationalism (and related positions on the mind, e.g., machine functionalism), to be a "stunning mistake."[6] Dennett has recently claimed that it is *Searle* who has made a stunning mistake: his claim is specifically that Searle's inference from RM's central zombie thought-experiment is obviously flawed, and fatally so. But, as we'll soon see, the argument based upon this thought-experiment is not only competent: once formalized, it becomes transparently valid. Moreover, the Searlean zombie argument can easily withstand Dennett's recent computationalist *Consciousness Explained* (= CE),[7] the achilles heel of which, interestingly, would appear to be precisely its vulnerability to zombie thought-experiments.

These thought-experiments arise from a situation lifted directly out of the toolbox most philosophers of mind, today, carry with them on the job: Your brain starts to deteriorate and the doctors replace it, piecemeal, with silicon chip workalikes, until there is only silicon inside your refurbished cranium.[8] Searle claims that at least three distinct possibilities arise from this gedanken-experiment:

[V1] The Smooth-as-Silk Variation: The complete silicon replacement of your flesh-and-blood brain works like a charm: same mental life, same sensorimotor capacities, etc.

[V2] The Zombie Variation: "As the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are

---

6 The Dennett-Searle clash has recently reached a new level of ferocity: Dennett claims that Searle is at best an exceedingly forgetful ([25], p. 203) philosopher:

> Is it possible that although Searle has at one time or another read all the literature, and understood it at the time, he has actually forgotten the subtle details, and (given his supreme self-confidence) not bothered to check his memory? For instance, has he simply forgotten that what he calls his reductio ad absurdum of my position (81 [in (Searle, 1992)]) is a version of an argument I myself composed and rebutted a dozen years ago? There is evidence of extreme forgetfulness right within the book. For instance...

In the next paragraph, speaking about another of Searle's supposed lapses, Dennett says, "But he forgets all this (apparently!) when forty pages later (107 [in (RM)]) he sets out to explain the evolutionary advantage of consciousness..." ([25]).

7 1991, Boston, Massachusetts: Little, Brown.

8 For example, the toolbox is opened and the silicon supplantation elegantly pulled out in Cole, D. and Foelber, R. (1984) "Contingent Materialism," *Pacific Philosophical Quarterly* **65.1**: 74–85.

indeed losing control of your external behavior...[You have become blind, but] you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me.'...We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same" ([52], 66–67).

[V3] The Curare Variation: Your body becomes paralyzed and the doctors, to your horror, give you up for dead.[9]

Searle wants to draw a certain conclusion from V2, the zombie variation, and it's this inference which turns Dennett nearly apoplectic. Here's a summary of the moral Searle wants to draw from V2, in his own words:

In [V2] we imagined that the mediating relationship between the mind and the behavior patterns was broken. In this case, the silicon chips did not duplicate the causal powers of the brain to produce conscious mental states, they only duplicated certain input-output functions of the brain. The underlying conscious mental life was left out ([52], 68).

And here is Dennett's reaction:

But that is only one of the logically possible interpretations of his second variation ... The other is the crucial one: while *you* ... are dying, *another* consciousness is taking over your body. The speech acts you faintly hear your body uttering are not yours, but they are also not nobody's! ... I cannot see how Searle could simply have overlooked this gaping loophole in his thought-experiment. But there it is ... I am baffled ([25], 198–99).

But what *exactly* does Searle want from V2? He tells us explicitly on page 69 of *The Rediscovery of the Mind* that he wants to establish via V2 and V3 that a certain trio of propositions is inconsistent. The trio, reproduced verbatim (p. 69):

(1) Brains cause conscious mental phenomena.

(2) There is some sort of conceptual or logical connection between conscious mental phenomena and external behavior.

(3) The capacity of the brain to cause consciousness is conceptually distinct from its capacity to cause motor behavior. A system could have consciousness without behavior and behavior without consciousness.

---

[9] This scenario would seem to resemble a real-life phenomenon: the so-called "Locked-In" Syndrome. See [43] (esp. the fascinating description on pages 24–25) for the medical details.

We can put things a bit more perspicuously, and put ourselves in position to assess the Dennett-Searle clash, if we represent the three propositions using elementary logical machinery: Bx iff x is a brain; Mx iff x causes (a full range of) mental phenomena; and Ex iff x causes (a full range of) external behavior. Then the trio, with Searle's underlying modal notions brought to the surface, and a denoting the brain of the character in our thought-experiments, becomes

(1*) $\exists x\ (Bx \wedge Mx)$

(2*) $\Box\forall x\ ((Bx \wedge Mx) \rightarrow Ex) \wedge \Box\forall x\ ((Bx \wedge Ex) \rightarrow Mx)$

(3*) $\Diamond\ (Ba \wedge Ma \wedge \neg Ea) \wedge \Diamond\ (Ba \wedge \mathrm{Ea} \wedge \neg Ma)$

The set {(1*), (2*), (3*)} is provably inconsistent, in garden variety contexts; the proof is trivial, for example, in quantificational S5 (which I happen to like) and the weaker T.[10] Dennett's objection, however, is that (3*) doesn't follow from V2. But this is hardly a gaping loophole; the situation is remedied merely by fine-tuning the zombie variation: Let $V2_1$ denote the one–(moribund)consciousness variation Searle describes, let $V2_2$ describe the two-consciousness variation Dennett describes (and, for that matter, let $V2_3$ denote the three-consciousness case, $V2_4$ the four, *ad infinitum*). Clearly, $\Diamond V2_1$ (as Dennett himself concedes in the quote above). And just as clearly this logical possibility implies the second conjunct of (3*) (and $\Diamond V3_1$ implies the *first* conjunct).

Now, Searle's ultimate aim is probably not to show {(1), (2), (3)} or its formal correlate inconsistent, for reaching this aim, as we have seen, is a matter of some pretty straightforward logic. Rather, Searle aims no doubt to refute the claim that there is a conceptual connection between conscious mentality and behavior, that is, he seeks to demonstrate the truth of (3*) and the falsity of (2*)—a result which follows when the inconsistency we have

---

[10] Systems like T and S5 can be determined by specifying certain rules of inference (which in both cases include the rules of first-order logic) and axiom-schemata. The key axiom-schema in T is the one known by that name, viz., $\Box\phi \rightarrow \phi$; the key axiom-schema in S5 is 5: $\Diamond\ \phi \rightarrow \Box\Diamond\phi$. (S5 includes as a theorem the interesting $\Diamond\Box\phi \rightarrow \Box\phi$, which becomes relevant later in the paper.) In both systems, moving a negation sign through a modal operator changes that operator (from diamond to box, and vice versa) in a manner perfectly analogous to the rule of quantifier negation in first-order logic. For a succinct presentation of the core ideas behind (propositional) S5 see Chapter 1 of [16] (a book which includes discussion of T and other systems as well). Here is how the proof goes. Proposition (1*) is superfluous. Then, e.g., instantiate appropriately on axiom- schema T to get, with (2*), by modus ponens, $\forall x\ ((Bx \wedge Mx) \rightarrow Ex)$; instantiate to $(Ba \wedge Ma) \rightarrow Ea)$, derive by propositional logic that $\neg((Ba \wedge Ma) \wedge \neg Ea)$, rewrite this by the rule known as necessitation to $\Box\neg$ ((Ba $\wedge$ Ma) $\wedge \neg$ Ea), and in turn rewrite this as $\neg\Diamond\neg\neg((Ba \wedge Ma) \wedge \neg Ea)$, and then, by double negation, as $\neg\Diamond((Ba \wedge Ma) \wedge \neg Ea)$, which of course contradicts (3*)'s first conjunct.

noted is combined with $\Diamond V2_1$, $\Diamond V3_1$ and $((\Diamond V2_1 \wedge \Diamond V3_1) \rightarrow (3^*))$.[11] Hereafter this argument is denoted by 'A_1'.

By this point the reader has doubtless realized that there is an opportunity for careful exegesis before us. In conceding the logical possibility of $V2_1$, Dennett does seem to grant all that Searle needs from the case. But must not Dennett somehow see Searle's aim differently? After all, why does he think it's crucial that the possibilities listed by Searle are *exhaustive*? My objective herein is not to explain, or explain away, Dennett's apparent lapse; my aim is to overthrow computationalism. Accordingly, I am happy to have arrived at A_1, and in the next section I proceed without further ado to adapt this argument to one specifically targeting this theory of mind—after which I offer a sustained defense of both the adaptation and A_1. However, in the course of this defense I cite and develop *seven* possible responses from Dennett, including one recently supplied by him through direct communication. These responses provide ample material for attempting the exegesis in question, and though for each one I will offer suggestions for how it can anchor the exegesis, I will leave detailed attempts to readers more concerned with hermeneutics than with whether or not computationalism is misguided.

2 Targeting Computationalism

It's easy enough to refine and then adapt what I have called Dennett's Dilemma so that it targets computationalism.

The first refinement is to replace talk of 'mental phenomena' with something more specific: I have in mind what is sometimes called **phenomenal consciousness**. Ned Block, in a recent essay on consciousness in *Behavioral and Brain Sciences* [4], calls this brand of consciousness **P-consciousness**. Here is part of his explication:[12]

---

11 For textual evidence that this is indeed Searle's goal, see p. 69 of RM.

12 Block distinguishes between P-consciousness and **A-consciousness**; the latter concept is characterized as follows:

> A state is access-conscious (A-conscious) if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous, i.e., poised to be used as a premise in reasoning, and (2) poised for [rational] control of action and (3) poised for rational control of speech. ([4], p. 231)

As I have explained elsewhere [10], it's plausible to regard certain extant, mundane computational artifacts to be bearers of A-consciousness. For example, theorem provers with natural language generation capability, and perhaps *any* implemented computer program (and therefore no doubt Pollock's OSCAR), would seem to qualify. It follows that a zombie would be *A*-conscious. In [10] I argue that because (to put it mildly here) it is odd to count (say) ordinary laptop computers running run-of-the-mill PASCAL programs as conscious in any sense of the term, 'A-consciousness' ought to be supplanted by suitably configured terms from its Blockian definition.

So how should we point to P-consciousness? Well, one way is via rough synonyms. As I said, P-consciousness is experience. P-conscious properties are experiential properties. P-conscious states are experiential states, that is, a state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are "what it is like" to have it. Moving from synonyms to examples, we have P-conscious states when we see, hear, smell, taste and have pains. P-conscious properties include the experiential properties of sensations, feelings and perceptions, but I would also include thoughts, wants and emotions. ([4], p. 230)

With the notion of P-consciousness in hand, and the "cognition is computation" core of computationalism in mind, it's easy to modify Searle's (1)–(3) so as to produce a parallel trio:[13]

(1_C) Persons are material things, viz., their brains (or some proper part of their central nervous systems).

(2_C) There is a conceptual or logical connection between P-consciousness and the structure of, and information flow in, brains: viz., Necessarily, if a person *a*'s brain instantiates a computation *c* (which must of course be of a particular type) from t_i to t_k of some Turing Machine (or other equivalent computational system) *m*, then person *a* enjoys a stretch of P-consciousness—from t_i to t_k—which is identical to *c*.

(3_C) A person's having P-consciousness is conceptually distinct from that person's brain being instantiated by a Turing Machine running through some computation.

Next, we can again employ some simple modal logic to formalize (1_C)–(3_C) in order to produce an inconsistent trio (1^*_C)–(3^*_C) that serves as a counterpart for (1*)–(3*).[14] The next move is to adjust $V2_I$ by adding the stipulation to the premise behind this thought-experiment that after Smith's brain begins to deteriorate, the doctors replace it, piecemeal, with silicon chip workalikes *which perfectly preserve the structure of, and computational flow in, that brain.* Call this new thought-experiment $V2^{TM}_I$. Finally, it's easy to use the inconsistency to fashion from A_I and $\Diamond V2^{TM}_I \rightarrow (3_C) \wedge (3^*_C)$ a parallel argument—call it A^C_I—the conclusion of which is the denial of (2_C), the heart of computationalism.

---

13 For a more formal version of (2_C) see my [7].

14 I leave the formalization to motivated readers. One way to go is to invoke a sorted calculus with *a*, *a'* ... ranging over persons, *c*, *c'* ... over computations, *s*, *s'* ... over stretches of consciousness, and $(t_i–t_k)$, $(t_i–t_k)'$ ... over intervals of time. Then if *Cxyz* is a predicate meaning that *x* enjoys *y* over *z*, (2_C) would start with $\Box\forall\, a\, \forall\, s\, \forall\, (t_i–t_k)\, Cas(t_i–t_k) \rightarrow$.

It follows that if Dennett's Dilemma cannot be escaped, (2_C) is overthrown, which in turn serves to overthrow computationalism itself.[15] I turn now to the task of closing off all possible escapes.

3 Can Dennett Dodge His Dilemma?

What would Dennett have to say for himself? It may be thought that Dennett need but point out that Searle's (2) claims only that "there is *some* sort of conceptual or logical connection between conscious mental phenomena and external behavior," where the italics are supplied by Dennett. For Dennett might then appeal to versions of functionalism wherein the link between mind and behavior isn't as strong as that implied by the modal (2*). For example, one brand of functionalism holds that what makes a mental state a state of a given type is the causal functional role it *typically* plays within an interconnected network of inputs, outputs and other states of the system. On this view, a given state can be of a specific type even if it fails to play the role typically played by such states, and even if it fails to result in any appropriately related behavior in the specific case. So this view provides an instantiation of the phrase 'some sort of conceptual connection,' and hence an instantiation of (2), but this instantiation isn't formalizable as (2*).

Unfortunately, Dennett would not succeed with such a move, for at least two reasons.

First, Searle would certainly be content to refute traditional brands of functionalism—brands including a modal conditional to the effect that if an organism o is in a certain compu-causal state s, then o is necessarily the bearer of a certain mental state s_m. In connection with this observation, it is important to note that the target of my adaptation of Searle is none other than a specification of such a modal conditional: (2_C). And that such a conditional be taken to capture the heart of computationalism is quite in keeping with the literature (e.g., [38], [3], [31], [56], [57], [42], [37], [39], [40], [21], [14], [54], [36]), which takes computation to reflect the *essence* of thinking. The idea is that thinking *is* computing, not that computing can be so configured as to produce a thing that seems to think but really doesn't (as in a zombie). Here is how Haugeland puts it:

> What are minds? What is thinking? What sets people apart, in all the known universe? Such questions have tantalized philosophers for millennia, but ... scant progress could be claimed ... until recently. For the current generation has seen a sudden and brilliant flowering in the philosophy/science of the mind; by now not only psychology but also a host of related disciplines are in the throes of a great intellectual revolution. And the epitome of the entire drama is *Artificial Intelligence*, the exciting new effort to make computers think. The fundamental goal of this research is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: *machines with minds*, in the full and literal

[15] Note that Pollock, in *How to Build a Person* [45], attempts to build the foundation for person building by first trying to establish (1_C) and (2_C).

sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, *computers ourselves* ([37], p. 2).

As many readers will remember, functionalist views taking the form of modal conditionals like (2_C) have been the target of "arbitrary realization" arguments, which involve thought-experiments designed to show the logical possibility of an organism instantiating compu-causal state *s* but failing to have *any* mental states.[16] Searle's zombie scenarios could be understood as thought-experiments intended to play the same role as those at the core of arbitrary realization arguments: (2*)'s second conjunct would be a formalization of the sort of modal conditional cited above, and $\Diamond V2_1$ would destroy this conjunct.[17]

This is not at all to say that V2 is, or can be distilled to, an arbitrary realization argument. In the tragic story of the poor soul in V2 who fades away when his brain is gradually replaced with silicon workalikes, there is nothing ridiculous: the workalikes are not tiny beer cans connected with string, for example. It's also important to distinguish the zombie in V2 from Swampman.[18] The Swampman has a more violent history than our zombie (and, as Bill Rapaport recently pointed out to me, history is thought by many to be crucial in these cases; cf. [41]), and it would be very difficult to flesh out the Swampman case with neuroscientific details (such fleshing out becomes important later in the paper).

The second reason why the move under consideration on Dennett's behalf—reading (2) so that it makes no claim about an intimate connection between external behavior and internal mentality—is not available to him is that though many philosophers these days would resist the view that you can *define* the having of a type of mental state exhaustively in terms of external

---

16 I have recently devised such thought-experiments to refute new, ingenious versions of machine functionalism explicitly designed to resist older, more primitive thought-experiments of the same general type. The purportedly inoculated brands of functionalism are specified by John Pollock [45]; my gedanken-experiments can be found in "Chapter VI: Arbitrary Realization" of my *What Robots Can and Can't Be* [14]. The older thought-experiments are due to Ned Block [5] and, ironically enough, John Searle [53].

17 After all, to say, as we have said on Dennett's behalf, that (2) can be read as "...there is only *some* sort of conceptual or logical connection between conscious mental phenomena and external behavior..." conveniently overlooks the fact that the modal (2*) is designed specifically to capture the notion that the connection in question *is*, whatever its specifics, conceptual/logical.

18 The Swampman was introduced by Donald Davidson:

> Suppose lightning strikes a dead tree in a swamp; I am standing nearby. My body is reduced to its elements, while entirely by coincidence (and out of different molecules) the tree is turned into my physical replica. My replica, The Swampman, moves exactly as I did; according to its nature it departs the swamp, encounters and seems to recognize my friends, and appears to return their greetings in English. It moves into my house and seems to write articles on radical interpretation. No one can tell the difference. ([17], 441)

behavioral conditions, Dennett would not, given his well-known "intentional stance" theory. This is readily confirmed by looking to some of the relevant writings. For example, consider Dennett's *Brainstorms* [27]. When at the start of that book Dennett introduces his notion of an *intentional system*, and the intentional stance built upon it, he makes clear that he rejects machine functionalism of the sort set out in our (2_C) (cf. [27], p. xvi). His version of this doctrine is laid out (on the same page) as follows.

(4) $\forall x$ (x believes that snow is white $\equiv x$ "realizes" some Turing machine k in logical state A)

In Chapter 2 of *Brainstorms* Dennett gives his argument against (4), which is based on a thought-experiment featuring two different face-recognition systems, each designed by a different team of AI engineers. The crux is that while both systems are "well characterized as believing that p" ([27], p. 26; p ranges over such things as "I've seen this face before"), by hypothesis they realize different Turing Machines in different states. Proposition (4) is thus supposed to be overthrown by the time-honored scheme: there is some scenario wherein its antecedent is true while its consequent is false. Whether or not Dennett succeeds in refuting (4), given our purposes, is immaterial. What matters is that Dennett's attack has been mounted from his armchair; his blow against (4) comes from a *thought*-experiment. This implies that (4) is to be read as asserting a *principled* connection between belief and Turing Machine computation. This in turn implies that (4) is a modal conditional of some sort, a soulmate for our (2_C). The equivalence in (4) is *logical* equivalence, not simply the material biconditional. For consider the following material biconditional.

(5) $\forall x$ (x is an American politician $\leftrightarrow x$ is corrupt)

This proposition cannot be overthrown by armchair reflection. The cynic who affirms it will not be obliged to change his attitude upon hearing that a philosopher has cooked up a coherent story about a virtuous Senator.

Dennett encapsulates his intentional stance via a conditional having precisely the form of (4), viz.,

(6) $\forall x$ (x believes that snow is white $\equiv x$ can be predictively attributed the belief that snow is white) (p. xvii, [27])

And (6) cannot be affirmed while at the same time the connection between mentality and behavior is said—via the rebuttal we are considering on Dennett's behalf—to be a non-conceptual/non-logical one.[19]

[19] Someone might say that I have here pressed a false dichotomy—because there is a third, "middle ground" conditional available to Dennett: one according to which a material

How else might Dennett try to dodge his dilemma? By carefully analyzing CE, and by ultimately asking Dennett himself, we can isolate and reject the remaining candidate responses:

3.1 Dennett's Objection From Method

Dennett might remind us that his complaint is about Searle's *method.* Well, true, Dennett strenuously objects in CE to Searle's emphasis on first-person introspection, which he regards to be a benighted vestige of Cartesian folk-psychology. Dennett would supplant Searle's method with his own, a method called "heterophenomenology"—but the problem is, heterophenomenology is neutral on the possibility of zombies! Dennett is quite explicit about this:

> [W]hat about the zombie problem? Very simply, heterophenomenology by itself cannot distinguish between zombies and real, conscious people, and hence does not claim to solve the zombie problem or dismiss it. ([26], 95)

It may be worth pointing out that Dennett's complaint about method, when applied to RM's central thought-experiment, seems aimed at a straw man: Propositions (1)–(3), as well as their formal counterparts (1*)–(3*), are apparently third-personish.[20] And, surely Dennett can't be saying that Searle's view is that we can establish $\Diamond\ \phi$, for any and all ϕ, only if ϕ can be conceived and "experienced" via the interior, "What-does-it-feel-like-to-me-on-the-inside-?" first-person point of view (since, e.g., Searle would hold, with us all, that it's logically possible that the Brooklyn Bridge turn instantly to jello, and would hold this in the absence of any phenomenological gymnastics).

3.2 Dennett's "Oops" Objection: Zombies vs. Zimboes

Faced with what our analysis has uncovered, Dennett might say that (oops) in his quoted reaction above he meant by the phrase "logically possible interpretation" not "account which describes a logical possibility," which is my reading, and surely the natural one, but something weaker like "candidate interpretation." This would be a rather desperate move. The problem with it, of course, is that not only is it (obviously!) logically possible that someone *offer* $V2_1$, but $V2_1$ *is* itself logically possible.[21] After all, Searle could, at the drop of a hat, provide a luxurious novel-length account of the scenario in question (or he could hire someone with the talents of a Kafka to do the job

---

conditional is said to be not logically necessary, but *physically* necessary. I refute this move separately in section 4.

20 Kieron O'Hara has pointed out to me that *Mx* is "available" only to the first-person, hence my insertion of the qualifier 'apparently' in the preceding sentence. Nothing substantive hinges on the claim that (1)–(3) are third-personish.

21 As nearly all those who write on the zombie topic agree. See, for example, Dretske [28], Block [4], Flanagan, [33], and Harnad [35].

for him).[22] Besides, if Dennett seriously maintains $\neg\Diamond V2_1$, where is the contradiction that the zombie scenario must then entail?

It might be said that I have moved hastily in the preceding paragraph, especially in light of another zombie-relevant section in Dennett's CE. More specifically, it might be said that Dennett not only might *say* that $\neg\Diamond V2_1$, he *does* say this in CE, and not only that: he produces an *argument* in CE for this position. It does appear that there is such an argument (on which Dennett has recently placed his chips[23]); it spans pages 304–14 of CE, and begins with

> In a series of recent papers, the philosopher David Rosenthal ([46], [50], [49], [47], [48]) has analyzed [the] everyday concept of consciousness and its relation to our concepts of reporting and expressing. He uncovers some structural features we can put to good use. First, we can use his analysis to ... show how it discredits the idea of zombies... ([26], 304)

What is the relevant part of Rosenthal's position? The answer, courtesy of his [46], can be put in declarative form:

> Def 1 s is a conscious mental state at time t for agent $a =_{df} s$ is accompanied at t by a higher-order, noninferential, occurrent, assertoric thought s' for a that a is in s, where s' is conscious or unconscious.[24]

Def 1 is said to be the **higher-order theory** (HOT) of consciousness. What sorts of examples conform to HOT? Dennett focuses on the state *wanting to be fed.* On Rosenthal's view, this state *is* a conscious state—and the reason it

---

22 Despite having no such talents, I usually spend twenty minutes or so telling a relevant short story to students when I present zombies via V2. In this story, the doomed patient in V2—Robert—first experiences an unintended movement of his hand, which is interpreted by an onlooker at perfectly natural. After more bodily movements of this sort, an unwanted sentence comes out of Robert's mouth—and is interpreted by an interlocutor as communication from Robert. The story describes how this weird phenomenon intensifies ... and finally approaches Searle's "late stage" description.

23 In his recent "The Unimagined Preposterousness of Zombies" [23] Dennett says the argument from CE which we are about to examine shows that zombies are not really conceivable.

24 Def 1's time index (which ought, by the way, to be a *double* time index—but that's something that needn't detain us here) is necessary; this is so in light of thought-experiments like the following. Suppose (here, as I ask you to suppose again below) that while reading Tolstoy's *Anna Karenina* you experience the state *feeling for Levin's ambivalence toward Kitty.* Denote this state by s^*; and suppose that I have s^* at 3:05 pm sharp; and suppose also that I continue reading without interruption until 3:30 pm, at which time I put down the novel; and assume, further, that from 3:05:01—the moment at which Levin and Kitty temporarily recede from the narrative—to 3:30 I'm completely absorbed in the tragic romance between Anna and Count Vronsky. Now, if I report at 3:30:35 to a friend, as I sigh and think back now for the first time over the literary terrain I have passed, that I feel for Levin, are we to then say that at 3:30:35 s^*, by virtue of this report and the associated higher-order state targeting s^*, becomes a conscious state? If so, then we give me the power to change the past, something I cannot be given.

is is that it's the target of a higher-order thought, viz., the thought that I want to be fed. Rosenthal's Def 1, of course, leaves open the possibility that the higher-order thought can be itself unconscious.

How can Dennett be read as using Def 1 as a defense against the zombie attack (i.e. A_1 and A_2)? We can construct a counter-argument on Dennett's behalf which involves not only zombies, but also a variant that Dennett introduces to ward off zombies: zimboes. "A zimboe," Dennett says, "is a zombie that, as a result of self-monitoring, has internal (but unconscious) higher-order informational states that are about its other, lower-order informational states" ([25], 310). The corresponding argument is expressed by Dennett in rather desultory prose, but it can be charitably reconstructed from pages 310–11 of CE as a *reductio* aimed at Searle:[25]

A_2

(7)	$\Diamond V2_1$	supp.
(8)	If $\Diamond V2_1$, then zimboes are logically possible.	
(9)	If zimboes are logically possible, then Turing Test-passing zimboes are logically possible.	
∴ (10)	Turing Test-passing zimboes are logically possible.	7, 8, 9
(11)	Turing Test-passing zimboes are *not* logically possible.	
∴ (12)	$\neg\Diamond V2_1$	4, 7, 8

This argument is obviously formally valid (as can be seen by symbolizing it in the propositional calculus); premises (8) and (9) seem to me to be above reproach; (7) is an assumption for indirect proof; and (10) is an intermediate conclusion. This leaves (11); why is this proposition supposed to be true? This proposition follows from a supporting sub-argument, viz.,

A_3

(13) If Turing Test-passing zimboes are logically possible, then it's logically possible that, in response to queries from the judge in the Turing Test, a zimboe's lower-level states become the target of higher-level, noninferential, occurrent, assertoric states.

(14) Def 1: *s* is a conscious mental state at time *t* for agent *a* $=_{df}$ *s* is accompanied at *t* by a higher-order, non-inferential, occurrent, assertoric thought *s´* for *a* that *a* is in *s*, where *s´* is conscious or unconscious.

---

25 I assume readers to be familiar with Alan Turing's [59] famous "imitation game" test of computational consciousness, now known as the **Turing Test**.

∴ (15) If Turing Test-passing zimboes are logically possible, it's logically possible that a Turing Test-passing zimboe is conscious. 13, 14

(16) It's necessarily the case that a zimboe is *un*conscious.

∴ (11) Turing Test-passing zimboes are *not* logically possible. 15, 16

Here, again, the reasoning can be effortlessly formalized (in propositional modal logic) and thereby shown to be valid. I grant (13), because Dennett's view that the judge in the Turing Test will unavoidably catalyze self-monitoring via questions to contestants like "Why did you say, a few minutes back, that..." is quite plausible. This leaves proposition (16) and Def 1 itself as the only potentially vulnerable spots. But (16) is true because *by definition* zimboes are not conscious. So we are left having to evaluate, in earnest, Def 1—no small task, since this definition is an account David Rosenthal has assembled, refined, and defended over many years in a number of intelligent papers.

Readers inclined to affirm my zombie attack, and to sustain it in the face of Dennett's HOT-based objection, may think there is a short-cut that obviates having to grapple with HOT. Specifically, the idea might be that Dennett begs the question: that A_2+A_3, and indeed any argument against $\Diamond V2_1$ which has Def 1 for a premise, is a case of *petitio principii.*

This move, I concede, is at least initially promising. For it would seem that Def 1 automatically implies that zombies are impossible, because according to this definition an unconscious state targeted at a lower-level state immediately implies (by *modus ponens* right-to-left across the biconditional) that the lower-level state is conscious. Isn't this circular? After all, the zombie thought-experiment is designed to reveal that it's perfectly conceivable that a behaviorally complex zombie (one that can pass itself off in human discourse as conscious, and therefore *a fortiori* one which can excel in the Turing Test[26]) exist! Rosenthal's views are undeniably convenient for Dennett, but perhaps they are *too* convenient. One way to focus this complaint is perhaps to ask: Why should we accept (13)? The intuition behind this premise, after all, was the same as that behind

(8) If $\Diamond V2_1$, then zimboes are logically possible;

and this proposition was to incarnate the intuition that we have only to make minor changes in the original zombie thought-experiment in order to have it portray zimboes. But now suddenly we find that that which the original

[26] We are now talking about Harnad's [36] ingenious *Total* Turing Test, the passing of which requires not only human-level linguistic behavior but sensorimotor behavior as well.

zombie thought-experiment is specifically designed to exclude—namely, intention-bearing states—is surreptitiously stipulated to be inherent in the zimboe thought-experiment! Is Dennett's move like rejecting the claim that time travel is logically possible by taking as *premise* a view of space-time according to which time travel is impossible?

No. The charge of *petitio* fails, for two reasons. First, it conflates 'intentional state' with 'P-conscious state.' The original zombie gedanken-experiment (and the amended version designed to overthrow computationalism) insists that P-consciousness is absent; it *doesn't* insist that *intentional* states (which at least on HOT may or may not be P-conscious) are excluded. So the move against Dennett under consideration is like rejecting the claim that time travel is logically *im*possible by taking as *premise* a view of space-time according to which time travel is *possible*. While most will agree that Dennett does not prove much by simply affirming Rosenthal's HOT, we must concede that neither does one prove anything by denying it. What is needed, then, *is* a direct attack on HOT itself—something I now provide.

3.3 A Direct Attack on Rosenthal's HOT

At first glance, it seems that HOT is quickly killed off by taking note of everyday experiences in which one is in a conscious state that is *not* the target of any higher-order state. For example, consider the state, *s*, feeling for Levin's ambivalence toward Kitty*, experienced while reading about these two characters as they move through Tolstoy's immortal *Anna Karenina*. Suppose that this state is experienced by you, as reader, while you are completely "transported" by the narrative. In such a case—so the story goes—there are no higher-order states directed at *s**.

But such thought-experiments are hardly conclusive. Rosenthal, in response, simply bites the bullet and insists that, unbeknownst to us, there just *is* a conscious-making higher-order state directed at *s**. He would say: "Look, *of course* you're not aware of any conscious states directed at *s**. This is just what my HOT predicts! After all, the state directed at *s** doesn't have a state directed at it, so it's unconscious; and whatever is unconscious is by definition something you're not aware of."

I think this response only gives Rosenthal a temporary reprieve, for there are other, sharper thought-experiments: Let *s´´* be a paradigmatic P-conscious state, say *savoring a healthy spoonful of deep, rich chocolate ice cream*. Since *s´´* is P-conscious, "there is something it's like" to be in this state. As Rosenthal admits about states like this one:

> When [such a state as *s´´*] is conscious, there is something it's like for us to be in that state. When it's not conscious, we do not consciously experience any of its qualitative properties; so then there is nothing it's like for us to be in that state. How can we explain this difference? ...

How can being in an intentional state, of whatever sort, result in there being something it's like for one to be in a conscious sensory state? ([46], pp. 24–25)

My question exactly. And Rosenthal's answer? He tells us that there are "factors that help establish the correlation between having HOTs and there being something it's like for one to be in conscious sensory states" (p. 26, [46]). These factors, Rosenthal tells us, can be seen in the case of wine tasting:

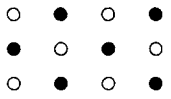
Learning new concepts for our experiences of the gustatory and olfactory properties of wines typically leads to our being conscious of more fine-grained differences among the qualities of our sensory states ... Somehow, the new concepts appear to generate new conscious sensory qualities. (p. 27, [46])

I confess I cannot help but regard Rosenthal's choice of wine tasting as tendentious. In wine tasting there is indeed a connection between HOTs and P-conscious states (the nature of which I don't pretend to grasp). But wine-tasting, as a source of P-consciousness, is unusually "intellectual," and Def 1 must cover all cases—including ones based on less cerebral activities. For example, consider fast downhill skiing. Someone who makes a rapid, "on-the-edge" run from peak to base will have enjoyed an explosion of P-consciousness; such an explosion, after all, will probably be the main reason such an athlete buys expensive equipment and expensive tickets, and braves the cold. But expert downhill skiers, while hurtling down the mountain, surely don't analyze the ins and outs of pole plants on hardpack versus packed powder surfaces, and the fine distinctions between carving a turn at 20 mph versus 27 mph. Fast skiers ski; they plunge down, turn, jump, soar, all at incredible speeds. Now is it really the case, as Def 1 implies, that the myriad P-conscious states $s_1, \ldots, s_n$ generated in a screaming top-to-bottom run are the result of higher-level, noninferential, assertoric, *occurrent* beliefs on the part of a skier k that k is in s_1, that k is in s_2, k is in s_4, ..., k is in s_n? Wine tasters do indeed sit around and say such things as that, "Hmm, I believe this Chardonnay has a bit of a grassy taste, no?" But what racer, streaking over near-ice at 50 mph, ponders thus: "Hmm, with these new parabolic skis, 3 millimeters thinner at the wait, the sensation of this turn is like turning a corner in a fine vintage Porsche." And who would claim that such thinking *results in* that which it's like to plummet downhill?

C F P Y
J M B X
S G R L

HOT is threatened by phenomena generated not only at ski areas, but in the laboratory as well. I have in mind an argument arising from the phenomenon known as **backward masking**. Using a tachistoscope,

psychologists are able to present subjects with a visual stimulus for periods of time on the order of milliseconds (one millisecond is 1/1000th of a second). If a subject is shown a 3 × 4 array of random letters (see the array above) for, say, 50 milliseconds (msecs), and is then asked to report the letters seen, accuracy of about 37% is the norm. In a set of very famous experiments conducted by Sperling [58], it was discovered that recall could be dramatically increased if a tone sounded after the visual stimulus. Subjects were told that a high tone indicated they should report the top row, a middle tone the middle row, and a low tone the bottom row. After the table above was shown for 50 msec, to be followed by the high tone, recall was 76% for the top row; the same result was obtained for the other two rows. It follows that a remarkable full 76% of the array is available to subjects after it appears. However, if the original visual stimulus is followed immediately thereafter by another *visual* stimulus in the same location (e.g., circles where the letters in the array appeared; see the array below), recall is abysmal; the second visual stimulus is said to backward mask the first. (See [2] for the seminal study.) Suppose, then, that a subject is flashed a series of visual patterns p_i, each of which appears for only 5 msec. In such a case, while there is something it is like for the subject to see p_i, it is very doubtful that this is because the subject thinks that she is in p_i. In fact, most models of human cognition on the table today hold that information about p_i never travels "far enough" to become even a potential object of any assertoric thought [1].



So, for these reasons, Def 1 looks to be massively implausible, and therefore the zombie attack has yet to be disarmed.[27]

3.4 Dennett's Objection from Racism

Dennett expresses this objection as follows:

> Notice, by the way, that this equivocation between two senses of "epiphenomenal" also infects the discussion of zombies. ['Zombie'] can be given a strong or weak interpretation, depending

---

[27] Interestingly enough, Dennett himself doesn't take Rosenthal's definition seriously. Nor, for that matter, does he feel, at this point in CE, any pressure to disarm the zombie attack:

> If ... Rosenthal's analysis of consciousness in terms of higher-order thoughts is rejected, then zombies can live on for another day's thought experiments. I offer this parable of the zimboes tongue in cheek, since I don't think either the concept of a zombie or the folk-psychological categories of higher-order thoughts can survive except as relics of a creed outworn. ([26], 313–14)

Since I do indeed reject Rosenthal's HOT, zombies live on, *by Dennett's own admission*, in the thought-experiments with which we began our investigation.

on how we treat ... indistinguishability to observers. If we were to declare that in principle, a zombie is indistinguishable from a conscious person, then we would be saying that genuine consciousness is epiphenomenal in the ridiculous sense. That is just silly. So we could say instead that consciousness might be epiphenomenal [in the sense that] although there was some way of distinguishing zombies from real people (who knows, maybe zombies have green brains), the difference doesn't show up as a functional difference to observers ... On this hypothesis, we would be able in principle to distinguish the inhabited bodies from the uninhabited bodies by checking for brain color. This is also silly, of course, and dangerously silly, for it echoes the sort of utterly unmotivated prejudices that have denied full personhood to people on the basis of the color of their skin. It is time to recognize the idea of the possibility of zombies for what it is: not a serious philosophical idea but a preposterous and ignoble relic of ancient prejudices. ([26], 405–6)

I have elsewhere [12] discussed the **penetrability** of zombies, where, put roughly, the idea is that zombies are penetrated if they are unmasked as empty-headed, if, to use Harnad's [35] phrase, there is found to be "nobody home." Penetrability seems to essentially correspond to what Dennett here calls "distinguishability;" so let's unite his property and mine under the predicate '*D*:' for a zombie *z*, Dz iff *z* is distinguished. In addition, let us invoke another predicate, *O*, which holds of a property *G* (we thus allow second-order constructs) just in case *G* is observed, and another property, *F*, which holds of a property if and only if that property "makes a functional difference." Finally, let's take explicit note of the fact that when Dennett refers to green brains, he must have in mind *any* property that could serve to distinguish zombies from "real people"—so that we can refer to any distinguishing property Δ. (No moderately intelligent defender of $\Diamond V2_1$ thinks that in the key thought-experiments the color of the "brains" involved, or any other simple property like this, is in any way relevant.) With this simple machinery in hand, we can charitably set out Dennett's argument in the previous quote, and we can connect it to the zombie attack I'm promoting in this paper:

A_4

(17) If Bringsjord's zombie attack on computationalism (A_1^C) is sound, then either $\Diamond\Box\forall z\neg Dz$ or $\forall z\exists\Delta\Diamond(O\Delta \wedge \neg F\Delta \wedge (O\Delta \rightarrow Dz))$

(18) It's not the case that $\Diamond\Box\forall z\neg Dz$.

(19) If $\forall z\exists\Delta\Diamond(O\Delta \wedge \neg F\Delta \wedge (O\Delta \rightarrow Dz))$, then the prejudices behind racism are acceptable.

(20) The prejudices behind racism are unacceptable.

$\therefore$ (21) It's not the case that $\forall z\exists\Delta\Diamond(O\Delta \wedge \neg F\Delta \wedge (O\Delta \rightarrow Dz))$. 19,20

$\therefore$ (22) Bringsjord's zombie attack on computationalism fails, i.e., A_1^C is unsound. 17,18,21

A_4 is of course formally valid. As to soundness, what of premises (18), (19), and (20)? The first of these is quite plausible; in fact, if the modal operators are interpreted in line with the modal system S5 visited above, we may even be able to give a little sub-proof of (18): $\Diamond\Box\phi \rightarrow \Box\phi$ is a (distinctive) validity in S5, so the first disjunct in (17)'s consequent, if we assume S5, becomes simply

$$\Box\forall z\neg Dz.$$

But this seems just plain wrong. Couldn't an observer always in principle be enlightened about "nobody homeness" by an oracle, or by God, or by other exotic but nonetheless coherent means? If so, then (18) is true.

But now what about (20)? This premise, I submit, is irreproachable, and who will contradict me? This leaves (19)—and here I think Dennett faces some rather rougher sledding. Why does he think this premise is true? Why does he think that affirming

$$(23) \qquad \forall z\exists\Delta\Diamond(O\Delta \wedge \neg F\Delta \wedge (O\Delta \rightarrow Dz))$$

entails an affirmation of racism? After all, isn't is somewhat implausible (not to mention harsh) to claim that those who affirm this formula are racists, or, in virtue of this affirmation, are at least willing to accept such prejudice? It's worth remembering, perhaps, that many of those who reject compu-causal functionalism will affirm (23). Are they thereby racist? I doubt it. It seems to me that everyone, Dennett included, must concede that racism, though perhaps aptly deemed a "cognitive sin," is usually understood to be a phenomenon a good deal less recherché than embracing the formula in question! Affirming (23) entails a certain attitude toward abstract properties in a thought-experiment; it does not entail a certain attitude toward *actual* beings. The quick way to grasp this is to simply change (23)'s quantification over zombies to (23′) quantification over a victimized race or group, and to then inquire if (23′) captures a racist attitude toward this group (it doesn't). The more rigorous route proceeds as follows.

First, note that (23)'s predicate O is ambiguous between "observed by a human" and "observed by an oracle (or a god ...)." As we noted when affirming (18), the second of these construals for O is the sort of thing we need to take seriously in our context. Let's use the subscripts $_h$ and $_o$ to disambiguate (23) into

$$(23_h) \qquad \forall z\exists\Delta\Diamond(O_h\Delta \wedge \neg F\Delta \wedge (O_h\Delta \rightarrow Dz))$$

$$(23_o) \qquad \forall z\exists\Delta\Diamond(O_o\Delta \wedge \neg F\Delta \wedge (O_o\Delta \rightarrow Dz))$$

Now, Dennett's premise (17) must allow both of these in the disjunction forming its consequent. To quickly make this change, let us stipulate that O

as it occurs in A_4 is a *disjunctive* property allowing for observation in either the O_h *or* O_o sense. It then becomes possible, I think, to show that (19) is false by devising a thought-experiment in which its antecedent is affirmed by those who are clearly not racists.[28]

Suppose that there are people—Oraks—living on the Earth-like planet Orak who believe that rocks, plants, individual molecules, grains of sand, electrons (and other sub-atomic particles) are not P-conscious. Suppose, furthermore, that Oraks have all thought long and hard about whether the coordinated movement of such objects changes anything: about whether such moving objects are P-conscious. Their conclusion, to a person, is that P-consciousness is still nowhere to be found. Let us imagine, in fact, that their conclusion stems from an affirmation of the arbitrary realization argument: they have considered scenarios in which, say, beer cans and string implement certain information-processing functions the Oraks have detected in their own brains, and have concluded that such contraptions would (obviously!, they say) lack P-consciousness. As to why Oraks *themselves* have P-consciousness, Oraks have a ready answer: the Oracle bestows a non-functional property upon their bodies, and *this* produces (or just is) P-consciousness. In fact, Oraks sometimes hear their Oracle speak: it thunderously says such things as, "Live!"—whereupon inert bodies glow for an instant and then move like those of healthy Oraks. Oraks, each and every one, affirm (23_o). Each and every Orak also lives a saintly life: they are loving, caring, altruistic, even-tempered, self-sacrificial, and so on. Finally,

---

[28] Though all this talk of oracles and gods may seem to be carrying us toward mysticism, there is actually an analogue for propositions of the form of (23_h) and (23_o) to be found in logic, specifically in the area of undecidable problems: It is well-known that there is no Turing Machine which can decide whether or not a fixed Turing Machine (or computer program, etc.) is ever going to halt. On the other hand, for every Turing Machine m, there is a fact of the matter as to whether or not m halts; m is either a halter or a non-halter (see [6] for elegant formal coverage of these matters). Now whereas it is generally thought to be impossible for a human to decide whether or not a Turing Machine halts (because, by computationalism itself, people don't have powers beyond Turing Machines), such is not the case for "oracles." Here, for example, is what we read in one of today's standard logicomathematical textbooks:

> Once one gets used to the fact that there are explicit problems, such as the halting problem, that have no algorithmic solution, one is led to consider questions such as the following. Suppose we were given a "black box" or, as one says, an *oracle*, which somehow can tell us whether a given Turing machine with given input eventually halts. Then it is natural to consider a kind of program that is allowed to ask questions of our oracle and to use the answers in its further computation. ([18], p. 197, emphasis his)

More specifically, if we reinterpret the predicate Dx as "decides whether or not x halts," and take the variable m to range over Turing Machines, the following two propositions are coherent and not implausible.

- $\forall m\, \Delta\Diamond(O_o\Delta \wedge \neg F\Delta \wedge (O_o\Delta \rightarrow Dm))$
- $\neg\forall m\, \Delta\Diamond(O_h\Delta \wedge \neg F\Delta \wedge (O_h\Delta \rightarrow Dm))$

every now and then the Oraks come upon a creature that looks like them, but which never bore the evanescent tell-tale glow, and cannot point to a time when the Oracle performed (what the Oraks call) ensoulment. These beings (known as "zombaks") the Oraks lovingly treat as they treat all other Oraks, but they nonetheless never reject (23_o).

This case seems perfectly coherent.[29] But then (19) is false, and A_4 fails.

Out of extreme charity, we might at this point imagine someone saying: "Sure the Oraks might be ever so nice to zombaks; and what is morally reprehensible about racism derives from unequal *treatment*—so in that sense they aren't racists. But there is also something intellectually reprehensible about reliance on an oracle to make a distinction when there is no way to see that distinction for oneself or to understand the mechanism by which another device (be it an oracle or something artificial) makes that distinction. We could quibble about whether or not it is appropriate to call this kind of intellectual defect 'racism,' but that would be to miss the point, I think, of Dennett's argument, which is that if you can't distinguish As from Bs except by pointing to features for which you have no explanation for the relevance of those features, then you are not being intellectually responsible."

This response only raises a red herring. The respondent admits that (19) is destroyed by my gedanken-experiment, so why does this new notion of "intellectual irresponsibility" matter? And why should Oraks be regarded intellectually irresponsible in the first place? After all, the Oraks *hear* the Oracle speak (and, for that matter, might interact with it in many and varied ways)! We can be more precise about the issue: Let f be some interesting function, one that defies classification in our world as (computationally) solvable or unsolvable. (There are many such functions.) Assume that in our world, despite decades of effort by first-rate mathematical minds, f stands as a mystery. Now suppose that f and the history of failed attempts to crack it are present in a world w_d in which both mere mortals (like ourselves) and a deity reside. Suppose, as well, that the mortal denizens of w_d are quite passionate about distinguishing between solvable problems (= As) and unsolvable problems (= Bs). If the deity classifies f as an A (B), why is it intellectually irresponsible for the mortals to agree? Since (by hypothesis) f is well-defined, either $f \in A$ or (exclusive) $f \in B$; moreover, there must be some determinate mathematical *reason* why f is in the set it's in. Why is it intellectually

---

[29] Note that this gedanken-experiment is *not* offered to establish that it's logically possible that it's logically possible that there are zombies. To sell the case as such would of course be to beg the question against certain opponents of my zombie attack, because $\Diamond\Diamond\phi \rightarrow \Diamond\phi$ is a validity in (e.g.) S5. What the case *does* show, it seems to me, is that in order to have racism one needs to have, in addition to an affirmation of something like (23), a corresponding prescription for how to think about and treat zombies or zombaks.

irresponsible to believe that a god has grasped this reason?[30] I conclude that (19) is indeed shot down by the story of the Oraks.

4 Two Final Moves

I see two remaining moves available to the computationalist in the face of A_I and A_I^C. The first is to make the claim that A_I^C can be dodged if one maintains not the denial of $\neg(2_C)$, but rather something like $(2_C')$, where this proposition is the result of changing the necessity operator in (2_C) to one of *physical* necessity.[31] This retreat is blocked by the brute fact that, as explained above, computationalists (e.g., [38], [3], [31], [56], [57], [42], [37], [39], [40], [21], [14], [54], [36]) have traditionally advanced the likes of the stronger (2).[32]

Besides, even those who suddenly decide to champion an idiosyncratic version of computationalism (according to which there is only a nomological connection between cognition and computation) will lose. This is because there would appear to be no reason why $V2_I$ and $V2_I^{TM}$ ought not to be regarded *physically* possible: Why couldn't a neuroscience-schooled Kafka write us a detailed, compelling account of $V2_I$ and $V2_I^{TM}$, replete with wonderfully fine-grained revelations about brain surgery and "neurochips"? Then, to generate the physics counterpart to A_I/A_I^C, we have only to change the modal operators to their physics correlates— $\Box$ to $\Box_p$ and $\Diamond$ to $\Diamond_p$, perhaps—and then invoke, say, some very plausible semantic account of this formalism suitably parasitic on the standard semantic account of logical modes.[33]

Note that the thought-experiment I have in mind, combined with such an account, does not merely establish that $\Diamond\Diamond_p V2_I^{TM}$. Such a proposition says that there is a possible world at which $V2_I^{TM}$ is physically possible—which is verified by imagining a possible world w in a cluster of worlds $w_1,\ldots,w_n$ comprising those which preserve the laws of nature in w, where $V2_I^{TM}$ is true not only at w, but at least one w_i. Let α be the actual world; let W_α^p denote the set of worlds preserving the laws of nature in α. The story I imagine Kafka telling scrupulously stays within W_α^p. Each and every inch of the

---

30 We could of course go on to flesh out the thought-experiment with additions like: The deity has in the past made many pronouncements on problems like *f*, and has in each case, after centuries of mortal effort secures on answer, been proved correct.

31 Proposition (2′) would say that it's *physically necessary* that a brain's causing mental phenomena implies corresponding external behavior.

32 By the way, the retreat would have for Dennett the welcome effect of trivializing Searle's pièce de résistance, for it implies that Searle's famous Chinese Room thought-experiment, designed to show that there is no *logical/conceptual* connection between symbol manipulation and mental phenomena, is trivial. Unfortunately, though Dennett isn't alone in wanting to dodge the Chinese Room, the view that the argument is trivial is an exceedingly solitary one. Note also that some have given versions of the Chinese Room designed from scratch to be physically possible. See, for example, "Chapter V: Searle," in [14].

33 For a number of such accounts, see [30].

thought-experiment is to be devised to preserve consistency with neuroscience and neurosurgery specifically, and biology and physics generally. My approach here is no different than the approach taken to establish that more mundane states of affairs are physically possible. For example, consider a story designed to establish that brain transplantation is physically possible (and not merely that it's logically possible that it's physically possible). Such a story might fix a protagonist whose spinal cord is deteriorating, and would proceed to include a step-by-step description of the surgery involved, each step described to avoid any inconsistency with neuroscience, neurosurgery, etc. It should be easy enough to convince someone, via such a story, that brain transplantation, at α, is physically possible. (It is of course much easier to convince someone that it's logically possible that it's physically possible that Jones' brain is transplanted: one could start by imagining (say) a world whose physical laws allow for body parts to be removed, isolated, and then made contiguous, whereupon the healing and reconstitution happens automatically, in a matter of minutes.)

Let me make it clear that I can easily do more than express my confidence in Kafka: I can provide an *argument* for $\Diamond V2_I^{TM}$ given that Kafka is suitably armed. There are two main components to this argument. The first is a slight modification of a point made recently by David Chalmers [15], namely, when some state of affairs ψ seems, by all accounts, to be perfectly coherent, the burden of proof is on those who would resist the claim that ψ is logically possible.[34] Specifically, those who would resist need to expose some contradiction or incoherence in ψ. I think most philosophers are inclined to agree with Chalmers here. But then the same principle would presumably hold with respect to *physical* possibility: that is, if by all accounts ψ seems physically possible, then the burden of proof is on those who would resist affirming $\Diamond_p\psi$ to indicate where physical laws are contravened.

The second component in my argument comes courtesy of the fact that $V2_I^{TM}$ can be modified to yield $V2_I^{NN}$, where the superscript 'NN' indicates that the new situation appeals not to Turing Machines, but to artificial neural networks, which are said to correspond to actual flesh-and-blood brains.[35]

---

[34] Chalmers gives the case of a mile-high unicycle, which certainly seems logically possible. The burden of proof would surely fall on the person claiming that such a thing is logically impossible. This may be the place to note that Chalmers considers it *obvious* that zombies are both logically and physically possible—though he doesn't think zombies are *naturally* possible. Though I disagree with this position, it would take us too far afield to consider my objections. By the way, Chalmers refutes ([15], 193–200) the only serious argument for the logical impossibility of zombies not covered in this paper, one due to Sydney Shoemaker [55].

[35] A quick encapsulation, given that while many readers are familiar with Turing Machines, less will be acquainted with artificial neural nets: Artificial neural nets (or as they are often simply called, 'neural nets') are composed of **units** or **nodes** designed to represent neurons, which are connected by **links** designed to represent dendrites, each of which has a numeric **weight**. It is usually assumed that some of the units work in symbiosis with

So what I have in mind for $V2_1^{NN}$ is this: Kafka really knows his stuff: he knows not only about natural neural nets, but also about artificial ones, and he tells us the sad story of Smith—who has his neurons and dendrites gradually replaced with artificial correlates in flawless, painstaking fashion, so that information flow in the biological substrate is perfectly preserved in the artificial substrate ... and yet, as in $V2_1^{TM}$, Smith's P-consciousness withers away to zero while observable behavior runs smoothly on. Now it certainly seems that $\Diamond_p V2_1^{NN}$; and hence by the principle we isolated above with Chalmers' help, the onus is on those who would resist $\Diamond_p V2_1^{NN}$. This would seem to be a *very* heavy burden. What physical laws are violated in the new story of Smith?

Some may retort that if the "physics version" of the zombie attack is sound, then beings with our behavioral repertoire, but without P-consciousness, could *in fact* have evolved in the actual world, on this very planet, under the constraints imposed by our laws of nature. Why then, they may go on to say, aren't *we* zombies? This question has already been eloquently raised in a slightly different form by those who merrily endorse $\Diamond V2_1$ [33]. I think the question can be cashed out in an explicit argument against either the core notion of P-consciousness or the claim that $\Diamond V2_1$. The gist of the argument is this:

Look, evolution implies that every significant mental property corresponds to some concrete behavioral "payoff," something that has survival value. But P-consciousness, in light of the arguments you promote, Bringsjord, corresponds to no such payoff. (The purported payoffs from P-consciousness can all be explained via information-processing mechanisms involved in

---

the external environment; these units form the sets of **input** and **output** units. Each unit has a current **activation level**, which is its output, and can compute, based on its inputs and weights on those inputs, its activation level at the next moment in time. This computation is entirely local: a unit takes account of but its neighbors in the net. This local computation is calculated in two stages. First, the **input function**, in_i, gives the weighted sum of the unit's input values, that is, the sum of the input activations multiplied by their weights:

$$in_i = \sum_j W_{ji} a_j.$$

In the second stage, the **activation function**, g, takes the input from the first stage as argument and generates the output, or activation level, a_i:

$$a_i = g(n_i) = g\left(\sum_j W_{ji} a_j.\right).$$

One common (and confessedly elementary) choice for the activation function (which usually governs all units in a given net) is the step function, which usually has a threshold t that sees to it that a 1 is output when the input is greater than t, and that 0 is output otherwise. This is supposed to be "brain-like" to some degree, given that 1 represents the firing of a pulse from a neuron through an axon, and 0 represents no firing. As you might imagine, there are many different kinds of neural nets. The main distinction is between **feed-forward** and **recurrent** nets. In feed-forward nets, as their name suggests, links move information in one direction, and there are no cycles; recurrent nets allow for cycling back, and can become rather complicated.

zombiehood; cf. [33].) Since evolution is of course true, it follows either that P-consciousness is a mirage, or $\neg\Diamond V2_1$.

There are at least two general ways to counter this argument. The first is to remember that evolution does allow for outright accidents, and to then point out that P-consciousness could be adventitious. (As Ned Block has recently pointed out to me, since at least all mammals are probably P-conscious, the accident would had to have happened quite a while ago.)

The second response, and the one I favor, is to step up to the challenge and show that certain behaviors *do* correspond to P-consciousness. I have elsewhere [11], [10] offered sustained arguments for the position that creativity, for example the creativity shown by a great dramatist, involves P-consciousness.[36] The second and final move open to the computationalist still bent on resisting my zombie attack is one Dennett has made in personal communication: concede $\Diamond V2_1$, and concede my corresponding arguments, but issue a reminder that zombies are not a "serious possibility."[37] In this regard zombies are said to be like gremlins in the internal combustion engine: it's logically possible that those little sedulous creatures are what make your car run, but no one takes this possibility seriously.[38] Unfortunately, this is just to change the subject. It's true that no one needing to fix an ailing Ford pops open the hood and reaches for bread crumbs to feed the gremlins; and it's also true that few indeed are those among us who wonder whether their friends are zombies. But such facts are perfectly consistent with each and every

---

36 Henrik Ibsen wrote:

> I have to have the character in mind through and through, I must penetrate into the last wrinkle of his soul. I always proceed from the individual; the stage setting, the dramatic ensemble, all that comes naturally and does not cause me any worry, as soon as I am certain of the individual in every aspect of his humanity. (reported in [32], p. xiv)

Ibsen's *modus operandi* is impossible for an agent incapable of P-consciousness. This is *not* to say that a zombie couldn't produce impressive text *without* using Ibsenian techniques.

37 There is a passage in CE consistent with this move. It is the last time Dennett discusses the zombie attack in his book:

> This book [argues] that if we are not urbane verificationists, we will end up tolerating all sorts of nonsense: epiphenomenalism, zombies, indistinguishable inverted spectra, conscious teddy bears, self-conscious spiders. ([26], 459)

The odd thing is that, as we have seen, Dennett nowhere in CE explains, let alone proves, that zombie thought-experiments, and the associated arguments, e.g., A_1, are nonsensical. Moreover, these experiments and arguments are part of the canonical arsenal used against verificationism!

38 Dennett claims that in light of this, zombies have become a public relations nightmare for philosophers, for when scientists hear that a major philosophical controversy boils down to zombies, they wear silly grins. But shouldn't we be concerned with constructing sound arguments and discarding unsound ones, regardless of how people *feel* about these arguments?

premise used above to refute the computational conception of mind. Besides, while the gremlin question promises to remain but an esoteric example dreamed up in an attempt to make a philosophical point, it is not hard to imagine a future in which the question of whether behaviorally sophisticated robots are or are not zombies is a pressing one.

5 Conclusion

Where does this leave us? Well, if computationalism is true, then (2_C), essentially the claim that appropriate computation suffices for P-consciousness, is as well. If argument A_1^C is sound, (2_C) falls, as does, by *modus tollens*, computationalism itself. A_1^C, as we have seen, can be formalized as a modal disproof patterned on A_1, the argument fashioned from the Dennett-Searle clash on zombie thought-experiments. (A_1 is based on gedanken-experiments designed to establish $\Diamond V2_1$ and $\Diamond V3_1$; A_1^C is based on one designed to establish $\Diamond V2_1^{TM}$.) A_1^C is formally valid. As to this argument's premises, they have emerged intact from a dialectical crucible fueled by the best objections computationalists can muster. Finally, attempts to tweak (2_C) so as to produce a view not targeted by A_1^C fails. (Here we considered zombie stories written (hypothetically) by Kafka to establish $\Diamond_p V2_1^{TM}$ and $\Diamond_p V2_1^{NN}$.) In the end, then, the zombie attack proves lethal: computationalism is dead.

References

[1] Ashcraft, M. H. (1994) *Human Memory and Cognition* (New York, New York: HarperCollins).

[2] Averbach, E. and Coriell, A. S. (1961) "Short-term Memory in Vision," *Bell System Technical Journal* **40**: 309–28.

[3] Barr, A. (1983) "Artificial Intelligence: Cognition as Computation," in Fritz Machlup, ed., *The Study of Information: Interdisciplinary Messages* (New York, New York: Wiley-Interscience), pp. 237–62.

[4] Block, N. (1995) "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences* **18**: 227–47.

[5] Block, N. (1980) "Troubles with Functionalism," in *Readings in Philosophy of Psychology Vol. I* (Cambridge, Massachusetts: Harvard University Press.)

[6] Boolos, G. S. and Jeffrey, R. C. (1989) *Computability and Logic* (Cambridge, UK: Cambridge University Press).

[7] Bringsjord, S. and Zenzen, M. (1997) "Cognition Is Not Computation: The Argument From Irreversibility," *Synthese* **113**: 285–320.

[8] Bringsjord, S. and Ferrucci, D. (forthcoming-a) *Artificial Intelligence and Literary Creativity Inside the Mind of Brutus, A Storytelling Machine* (Mahwah, New Jersey: Lawrence Erlbaum).

[9] Bringsjord, S. and Zenzen, M. (forthcoming-b) *Super-Minds: A Defense of Uncomputable Cognition* (Dordrecht, The Netherlands: Kluwer).

[10] Bringsjord, S. (1997) "Consciousness by the Lights of Logic and Common Sense," *Behavioral and Brain Sciences* **20.1**: 144–46.

[11] Bringsjord, S. (1995) "Pourquoi Hendrik Ibsen Est-Il Une Menace pour La Littérature Générée Par Ordinateur?" (traduit par Michel Lenoble) in *Littérature et Informatique la Littérature Générée Par Orinateur*, Alain Vuillemin, ed. (Arras, France: Artois Presses Université).

[12] Bringsjord, S. (1995) "In Defense of Impenetrable Zombies," *Journal of Consciousness Studies* **2.4**: 348–51.

[13] Bringsjord, S. (1995) "Could, How Could We Tell If, and Why Should–Androids Have Inner Lives," chapter in the *Android Epistemology* (Cambridge, Massachusetts: MIT Press), pp. 93–122. Ken Ford, Clark Glymour and Pat Hayes, editors.

[14] Bringsjord, S. (1992) *What Robots Can and Can't Be* (Dordrecht, The Netherlands: Kluwer).

[15] Chalmers, D. (1996) *The Conscious Mind: In Search of a Fundmental Theory* (Oxford, UK: Oxford University Press).

[16] Chellas, B. F. (1980) *Modal Logic: An Introduction* (Cambridge, UK: Cambridge University Press).

[17] Davidson, D. (1987) "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* **60**: 441–58.

[18] Davis, M. D., Sigal, R. and Weyuker, E. J. (1994) *Computability, Complexity, and Languages* (San Diego, California: Academic Press).

[19] Davis, W. (1988) *Passage of Darkness: The Ethnobiology of the Haitian Zombie* (Chapel Hill, North Carolina: University of North Carolina Press).

[20] Davis, W. (1985) *The Serpent and the Rainbow* (New York, New York: Simon & Shuster).

[21] Dietrich, E. (1990) "Computationalism," *Social Epistemology* **4.2**: 135–54.

[22] Dennett, D. C. (1996) "Cow-sharks, Magnets, and Swampman," *Mind and Language* **11.1**: 76–77.

[23] Dennett, D. C. (1995) "The Unimagined Preposterousness of Zombies," *Journal of Consciousness Studies* **2.4**: 322–26.

[24] Dennett, D. C. (1994) "The Practical Requirements for Making a Conscious Robot," *Philosophical Transactions of the Royal Society of London* **349**: 133–46.

[25] Dennett, D. C. (1993) "Review of Searle's The *Rediscovery of the Mind*" *Journal of Philosophy* **90.4**: 193–205.

[26] Dennett, D. C. (1991) *Consciousness Explained* (Boston, Massachusetts: Little, Brown).

[27] Dennett, D. C. (1978) *Brainstorms* (Cambridge, Massachusetts: MIT Press).
[28] Dretske, F. (1996) "Absent Qualia," *Mind and Language* **11.1**: 78–85.
[29] Dretske, F. (1995) *Naturalizing the Mind* (Cambridge, Massachusetts: MIT Press).
[30] Earman, J. (1986) *A Primer on Determinism* (Dordrecht, The Netherlands: D. Reidel).
[31] Fetzer, J. (1994) "Mental Algorithms: Are Minds Computational Systems?" *Pragmatics and Cognition* **2.1**: 1–29.
[32] Fjelde, R. (1965) Foreword in Ibsen, H. (1965) *Four Major Plays* (New York, New York: New American Library).
[33] Flanagan, O. and Polger, T. (1995) "Zombies and the Function of Consciousness," *Journal of Consciousness Studies* **2.4**: 313–21.
[34] Glymour, C. (1992) *Thinking Things Through* (Cambridge, Massachusetts: MIT Press.
[35] Harnad, S. (1995) "Why and How We Are Not Zombies," *Journal of Consciousness Studies* **1**: 164–67.
[36] Harnad, S. (1991) "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem," *Minds and Machines* **1.1**: 43–55.
[37] Haugeland, J. (1981) *Artificial Intelligence: The Very Idea* (Cambridge, Massachusetts: MIT Press).
[38] Hobbes, T. (1839) *De Corpore*, chap. 1, in *English Works*, ed. Molesworth, reprinted in (1962) *Body, Man and Citizen* (New York, New York: Collier).
[39] Hofstadter, D.R. (1985) "Waking Up from the Boolean Dream," Chapter 26 in his *Metamagical Themas: Questing for the Essence of Mind and Pattern* (New York, New York: Bantam), pp. 631–65.
[40] Johnson-Laird, P. (1988) *The Computer and the Mind* (Cambridge, Massachusetts: Harvard University Press).
[41] Millikan, R. G. (1996) "On Swampkinds," *Mind and Language* **11.1**: 103–17.
[42] Newell, A. (1980) "Physical Symbol Systems," *Cognitive Science* **4**: 135–83.
[43] Plum, F. and Posner, J.B. (1972) *The Diagnosis of Stupor and Coma* (Philadelphia, Pennsylvania: F. A. Davis).
[44] Pollock, J. (1995) *Cognitive Carpentry: A Blueprint for How to Build a Person* (Cambridge, Massachusetts: MIT Press).
[45] Pollock, J. (1989) *How to Build a Person: A Prolegomenon* (Cambridge, Massachusetts: Bradford Books, MIT Press).
[46] Rosenthal, D. M. (forthcoming) "State Consciousness and What It's Like," Title TBA (Oxford, UK: Clarendon Press).

[47] Rosenthal, D. M. (1990) "Why Are Verbally Expressed Thoughts Conscious?" ZIF Report No. 32, Zentrum für Interdisziplinäre Forschung, Bielefeld, Germany.

[48] Rosenthal, D. M. (1990) "A Theory of Consciousness," ZIF Report No. 40, Zentrum für Interdisziplinäre Forschung, Bielefeld, Germany.

[49] Rosenthal, D. M. (1989) "Thinking That One Thinks," ZIF Report No. 11, Research Group on Mind and Brain, Perspective in Theoretical Psychology and the Philosophy of Mind, Zentrum für Interdisziplinäre Forschung, Bielefeld, Germany.

[50] Rosenthal, D. M. (1986) "Two Concepts of Consciousness," *Philosophical Studies* **49**: 329–59.

[51] Russell, S. and Norvig, P. (1995) *Artificial Intelligence: A Modern Approach* (Englewood Cliffs, New Jersey: Prentice Hall).

[52] Searle, J. (1992) *The Rediscovery of the Mind* (Cambridge, Massachusetts: MIT Press).

[53] Searle, J. (1983) *Intentionality* (Cambridge, UK: Cambridge University Press).

[54] Searle, J. (1980) "Minds, Brains and Programs," *Behavioral and Brain Sciences* **3**: 417–24.

[55] Shoemaker, S. (1975) "Functionalism and Qualia," *Philosophical Studies* **27**: 291–315.

[56] Simon, H. (1980) "Cognitive Science: The Newest Science of the Artificial," *Cognitive Science* **4**: 33–56.

[57] Simon, H. (1981) "Study of Human Intelligence by Creating Artificial Intelligence," *American Scientist* **69.3**: 300–309.

[58] Sperling, G. (1960) "The Information Available in Brief Visual Presentations," *Psychological Monographs* **74**: 48.

[59] Turing, A. M. (1964) "Computing Machinery and Intelligence," in A. R. Andersen, ed., *Minds and Machines*, Contemporary Perspectives in Philosophy Series (Englewood Cliffs, New Jersey: Prentice Hall), pp. 4–30.