COGNITION IS NOT COMPUTATION: THE ARGUMENT FROM IRREVERSIBILITY*

Selmer Bringsjord
Dept. of Philosophy, Psychology & Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute
Troy NY 12180
selmer@rpi.edu • http://www.rpi.edu/∼brings

Michael Zenzen
Dept. of Philosophy, Psychology & Cognitive Science
Rensselaer Polytechnic Institute
Troy, NY 12180
zenzem@rpi.edu

November 14, 2000

Abstract

The dominant scientific and philosophical view of the mind — according to which, put starkly, cognition is computation — is refuted herein, via specification and defense of the following new argument: Computation is reversible; cognition isn't; ergo, cognition isn't computation. After presenting a sustained dialectic arising from this defense, we conclude with a brief preview of the view we would put in place of the cognition-is-computation doctrine.

*We are indebted to Bill Rapaport, Pat Hayes, Ken Ford, Marvin Minsky, Jim Fahey, two anonymous referees (who provided particularly insightful comments), and many Rensselaer students. These people provided trenchant objections which saw to the evolution of the present version from a rather inauspicious primogenitor.

1 Introduction

The dominant scientific and philosophical view of the mind — put starkly, that cognition is computation — is refuted herein, via specification and defense of the following new argument: Computation is reversible; cognition isn't; ergo, cognition isn't computation. The specification of the argument involves a quartet: (i) certain elementary theorems from computability theory, according to which computation is reversible; (ii) the doctrine of agent materialism, according to which, contrary to any sort of dualistic view, human agents (= human persons) are physical things whose psychological histories are physical processes; (iii) the introspection- and physics-supported fact that human cognition is *not* reversible; and (iv) the claim — fundamental to AI and Cognitive Science, and, again, put roughly for now — that cognition is computation. The basic structure of the argument is straightforward: the conjunction of (i), (ii) and (iii) entails the falsity of (iv).

Our plan is as follows. In Section 2 we take some preliminary steps toward unpacking the "cognition is computation" slogan. In Section 3 we provide a rudimentary but sufficient-for-present-purposes account of recursion-theoretic reversibility. In Section 4 we present the Argument From Irreversibility. Section 5 is a sustained dialectic arising from objections to the argument — a dialectic which generates specification of the argument presented in Section 4. In the final section, 6, we express our intuitions about what should supplant the view that cognition is computation, and by doing so offer a glimpse of our forthcoming monograph on "uncomputable cognition" [2b].

2 The Computational Conception of Mind

The view that cognition is computation needs little introduction. Propelled by the writings of innumerable thinkers (e.g., [19]; [2]; [16]; [36]; [37]; [28]; [18]; [20]; [23]; [14]; [2e]; [35]; [17] — and this touches but the tip of a mammoth iceberg of relevant writing), this view has reached every corner of, and indeed energizes the bulk of, contemporary Artificial Intelligence (AI) and Cognitive Science (Cog Sci). The view has also touched nearly every major college and university in the world; even the popular media have, on a global scale, preached the computational conception of mind. Of course, this conception is as protean as it is pandemic; the cognition-is-computation slogan competes for equal time with a number of others. For example, for

starters we have

- Thinking is computing.
- People are computers (perhaps with sensors and effectors).
- People are Turing machines (perhaps with sensors and effectors).
- People are finite state automata (perhaps with sensors and effectors).
- People are neural nets (perhaps with sensors and effectors).
- Cognition is the computation of Turing-computable functions.

There are differences, and in some cases significant differences, between these sorts of locutions. (We discuss some of these differences below.) But surely there is a great and undeniable (though confessedly vague) commonality in the works — a commonality captured, for example, by Haugeland:

> 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 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* ([18], p. 2).

This conveys the core spirit of "Strong" AI, which wavers not a bit in the face of questions about whether sensors and effectors are necessary, or where in the Chomsky Hierarchy from finite state automata to Turing Machines people fall. Nonetheless, it will facilitate matters if we have a rather more focussed version of the doctrine on the table. Accordingly, we will say, following [16] closely, that the cognition-is-computation view — sometimes called 'Computationalism,' sometimes 'Strong AI,' etc. — amounts to the following: People (or minds, or brains) are computers. Computers, in turn, are essentially Turing Machines (or other equivalent automata). Hence, the boundaries of computability define the boundaries of cognition. It's easy

enough to render this view wholly declarative, and hence quite unmistakable, using the predicate calculus, once the relevant predicates are defined by Mx iff x is a Turing Machine and Px iff x is a person. For simplicity, let's say:

Proposition 1. $\forall x\,(Px \Rightarrow \exists y\,(My \wedge x = y))$.

Later on we'll propose a more "fine-grained" version of Computationalism, one which makes explicit reference to both cognition and computation.[1] But before moving toward this version, one remark. There have of course been prior attacks on Computationalism (e.g., John Searle's [35] infamous Chinese Room Argument). To our knowledge, however, none of these attacks, with the possible exception of [26] and Chapter IX of [2e], have attempted to show that the falsity of this doctrine *follows from the very foundations upon which it's built.* In the case of Searle's Chinese Room Argument, everything hinges on an ingenious but undeniably fanciful thought-experiment. The Argument From Irreversibility is based, for the most part, on provable facets (e.g., reversibility) of computation which appear by the lights of both common-sense and elementary physics to be at odds with what we can come to know about cognition.

3 Rudiments of Computational Reversibility

Computationalism relies upon automata like Turing Machines (TMs) to fix the concept of computation. But what's a Turing Machine? To some readers these automata will be old hat, but let's play it safe and ask cognoscenti to return with us to the elementary aspects of computability theory upon which the Argument From Irreversibility is built.

Put intuitively, TMs include a **two-way infinite tape** divided into squares, a **read/write head** for writing and erasing **symbols** (from some

[1] But due to space constraints, nowhere in this paper do we provide a detailed, comprehensive account of Computationalism. Such an account would probably need to include careful versions of at least the following five propositions.

1. A function f is effectively computable if and only if f is Turing-computable (Church's Thesis).
2. Proposition 1.
3. The Turing Test is valid.
4. Computationalists will succeed in building persons.
5. Computationalists will succeed in building Turing Test-passing artifacts. (This proposition is presumably entailed by its predecessor.)

finite, fixed **alphabet**) on and off this tape, a **finite control unit** which at any step in a computation is in one particular state from among a finite number of possible states, and a set of **instructions** (= program) telling the machine what to do, depending upon what state it's in and what (if anything) is written on the square currently scanned by it's head. Formally speaking, a TM is a set, specifically a quintuple (S, Σ, s, h, f), where: S is a finite set (of states); Σ is an alphabet containing the special "blank" symbol b, but not the symbols L (for "left") and R (for "right"); $s \in S$ is the **initial state**; $h \notin S$ is the **halt state**; and f is a **transition function** from $S \times \Sigma$ to $(S \cup \{h\} \times (\Sigma \cup \{L, R\})$.

This formalism makes it easy to render the notion of a **configuration** of a TM precise,[2] but an informal account will suffice for our purposes: We'll say that a configuration is composed of four pieces of information: first, the state the TM is in; second, information representing what's to the left of the square currently scanned by the read/write head; third, the single symbol being scanned; and four, the contents of the tape to the right of the head. In a moment, we'll concretize the notion of a configuration by explaining and analyzing the particular TM specified in Figure 1. But first some notation for computations: Let $C_1, C_2, ...$ denote configurations. We write $C_i \vdash_M C_k$ to indicate that TM M has moved "in one step" from configuration C_i to C_k. For every TM M, $\vdash_M^*$ is the reflexive, transitive closure of $\vdash_M$. A **computation** by some TM M is a sequence of configurations $C_1, C_2, C_3, ..., C_n$, where $n \geq 1$ such that

$$C_1 \vdash_M C_2 \vdash_M C_3 \vdash_M \cdots \vdash_M C_n.$$

There are many ways to steamline the full set-theoretic description of TMs. One such method is the state diagram approach used in Figure 1. This TM, "Gordon's 19 in 186," is designed to start on a 0–filled infinite tape and produce, after 186 steps, 19 1's.[3]

[2]Formally, keeping in mind that after a while, in both directions, the tape is populated by infinitely many blanks, a configuration is a member of the set $S \cup \{h\} \times ...b \circ \Sigma^* \times \Sigma \times \Sigma^* \circ b...$

[3]To the uninitiated, this computation will doubtless sound remarkably unimpressive, but initiated ought to note that this machine is the result of Genetic Algorithm-powered search (engineered by Gordon Greene) in the (enormous) space of 6-state TMs for "Busy Beaver" candidates in the "quaduple" formalism. (Currently, the most productive known 6-state machine can produce 21 1's. The machine — built by Chris Nielsen — can be seen (in flow graph form) and obtained by linking through Bringsjord's web site (URL above), under the course *Symbolic Logic*. Alternatively, go directly to http://csli-

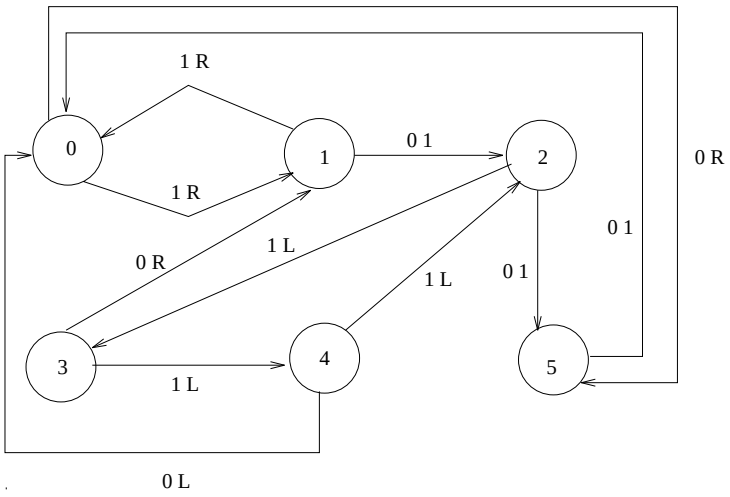


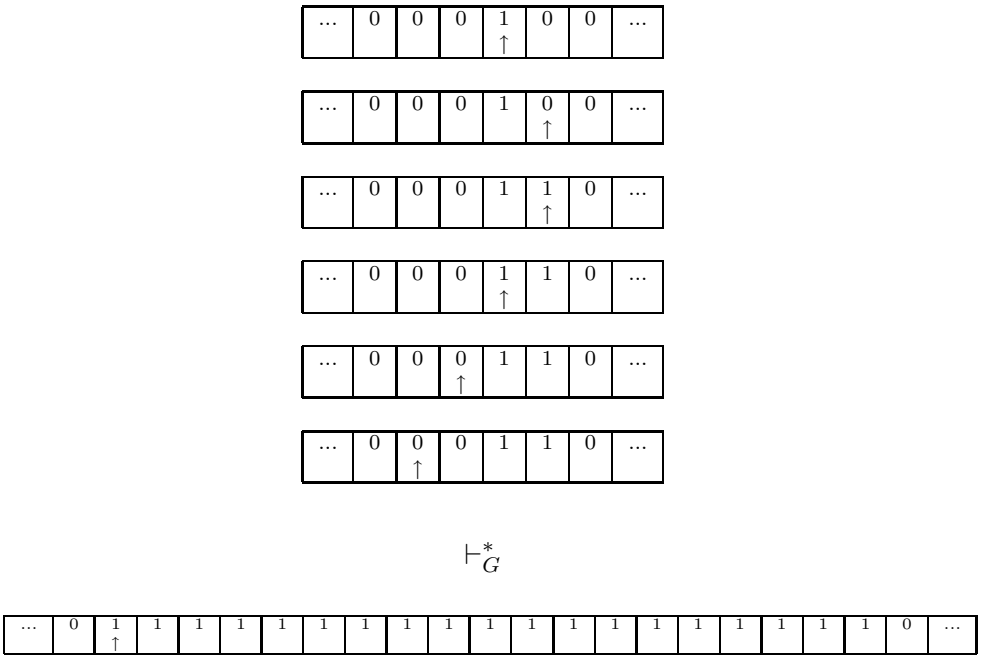
Figure 1: Gordon's 19 in 186

Let's "hand simulate" an initial segment of the computation of Gordon's TM — let's label the machine G — so that we completely fix the core mathematical concepts. The alphabet used is simply $\{0, 1\}$. The initial state of G is 0 (represented by the node labeled 0), and at the outset we'll assume that the tape is filled with 0's. The first thing G does is check to see what symbol it finds under its read/write head. In this case it initially finds a 0, so the arc labeled with 0 R is taken, which means that the head moves one square to the right and the machine enters state 5. At this point, since there is another 0 found beneath the head, the 0 is changed to a 1, and the machine reenters state 0. It now finds a 1, and hence takes the arc labeled 1 R to state 1 (i.e., the machine moves its head one square to the right, and then enters state 1) — etc. The machine's activity can be perfectly captured by a tedious catalogue of its configurations from start to finish, e.g.,

...	0	0	0 ↑	0	0	0	...

...	0	0	0	0 ↑	0	0	...

www.stanford.edu/hp/Beaver.html.) In the "Busy Beaver" function, $f_{bb} : \mathbf{N} \longrightarrow \mathbf{N}$ ($\mathbf{N}$ here denotes the natural numbers), $f_{bb}(n)$ yields the greatest number of 1's an n-state TM, starting on a black tape, can leave after halting. This function, and the corresponding proof that it's uncomputable, is due to Rado [32], who used $\Sigma(n)$ for f_{bb}. Busy Beaver candidates are those n-state TMs which appear to produce $f_{bb}(n)$ 1's.



If this is your first exposure to TMs, you will doubtless be struck by how primitive and unassuming they are. But the surprising thing is that TMs apparently capture computation *in all its guises*. More precisely, whatever can be accomplished by way of an algorithm, by way of a programmed supercomputer, by way of a neural network, a cellular automaton, etc. — whatever can be accomplished by any of these can be accomplished by a TM.[4] Furthermore, we know that augmenting the architecture our TMs doesn't give them any additional power. For example, if we give a TM *two* tapes rather than one, nothing which was impossible for the one-tape machine becomes doable for the two-tape creature. (This is a fact we revisit below.) Usually, such a fact is established through a so-called "simulation proof," the essence of which consists in showing that the new automaton can be perfectly simulated by the original one.[5]

Now, Bennett [6] has established that TM computation can be logically reversed in a "useful" way. It was clear before Bennett that computation can

[4]See [2f] for a discussion of the consequences of this fact for AI.

[5]The interested reader can consult an octet of books we find useful: For broad coverage of the basic material, see [24], [15], [9], and [21]. For a nice comprehensive discussion of computability theory that includes succinct coverage of uncomputability, including the Arithmetic Hierarchy, see [10] and the difficult but rewarding [40]. [29] contains a very nice discussion of the Chomsky Hierarchy. And, of course, there's always the classic [33].

be logically reversed in *non*-useful ways, and in fact this result is sufficient for our argument. Non-useful ways of reversing TM computation are easy to come by; here's one: First, go get some paper, a pencil, and an eraser. Next, watch a given TM M in action, recording each configuration through which it passes. Finally, using your list of configurations, build a new TM M_R which goes from the last entry on your list to the first in such a way that it visits your entries in reverse sequence.

Though we are informal here, it should be obvious that (in keeping with the inaugural writings of Turing and Post, in which the notion of a human computist is primitive) we have here a straightforward *algorithm*, call it $\mathcal{A}$, for reversing computation. You are encouraged to try your hand at simulating $\mathcal{A}$ on the TM shown in Figure 1. All you need is some patience, and, as we say, a pencil, an eraser, and sufficient paper.

If you're short of time, we suggest a shortcut constructive "proof" as a substitute: Purchase the software "Turing's World"^{TM},[6] run one of the TMs that come bundled with this package, and then use the REVERSE command to step back through the configurations your selected machine has run. You can be sure that the existence of some such algorithm as $\mathcal{A}$ is precisely what enables you to carry out this reversal using "Turing's World"^{TM}.

It will be helpful if we set off the relevant proposition as a theorem:[7]

Theorem 1. For every computation $C_1 \vdash_M C_2 \vdash_M C_3 \vdash_M \cdots \vdash_M C_n, n \in \mathbf{Z}^+$, there exists a computation $C_n \vdash_{M'} C_{n-1} \vdash_{M'} C_{n-2} \vdash_{M'} \cdots \vdash_{M'} C_1$ which can be obtained from the original computation via some algorithm $\mathcal{A}$.

4 The Argument From Irreversibility

4.1 The Starting Point: Proposition 1

We begin by refining Proposition 1 so that it clarifies and narrows the "cognition is computation" slogan:

[6] The "Turing's World"^{TM} software comes with [3].

[7] Note that by Church's Thesis Theorem 1 is interchangeable with this theorem (which is easy to prove *without* CT):

Theorem 1′. For every computation $C_1 \vdash_M C_2 \vdash_M C_3 \vdash_M \cdots \vdash_M C_n, n \in \mathbf{Z}^+$, there exists a computation $C_n \vdash_{M'} C_{n-1} \vdash_{M'} C_{n-2} \vdash_{M'} \cdots \vdash_{M'} C_1$ which can be obtained from the original computation via some TM M^*.

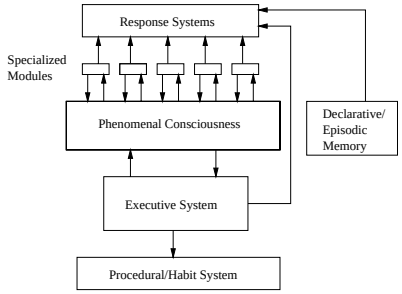


Figure 2: Schacter's Model

Proposition 1′. $\forall x\,(Px \wedge x$ is conscious from t_i to $t_{i+k} \Rightarrow \exists y\,(My \wedge x = y \wedge C_j \vdash_y C_{j+1} \vdash_y \cdots \vdash_y C_{j+p}))$, where this computation is identical to the consciousness x enjoys through $[t_i, t_{i+k}]$.

4.2 Clarifying 'Consciousness'

Note that Proposition 1′ marks a shift from talk of cognition to talk of consciousness. 'Consciousness' is a narrower term than 'cognition.' Cognition, that is, includes consciousness.[8] It follows that if consciousness isn't reversible, then the broader phenomenon of cognition isn't either. (Of course, many elements of cognition *are* computable.)

It's important to realize that we have something rather specific in mind when we use the term 'consciousness' in this paper: we have in mind what is sometimes called **phenomenal consciousness**. Ned Block [8], in a recent essay on consciousness in *Behavioral and Brain Sciences*, calls this brand of consciousness **P-consciousness**. Here's part of his explication of this concept:

> So how should we point to P-consciousness? Well, one way is via rough synonyms. As I said, P-consciousness is experience. P-conscious prop-

[8]This is borne out by, among other things, looking at proposed comprehensive models of cognition in cognitive science: these models often include some component claimed to account for consciousness. For example, Schacter [34] gives us the the picture of cognition shown in Figure 2. For a survey of models like this one, see [1]; for a penetrating analysis of such models (including Schacter's) see [8]. For those interested in charting the first-order formalization of our argument, the relevant sentence in first-order logic would be a full symbolization of $\forall x(x$ is cognizing $\rightarrow x$ is conscious), where 'cognizing' is understood to indicate the full scope of human cognition as purportedly captured in models like Schacter's.

> erties 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. ([8], p. 230)

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. ([8], p. 231)

Note that it's plausible to regard certain computational artifacts to be bearers of A-consciousness (e.g., theorem provers with natural language generation capability), whereas we shall now attempt to establish that P-consciousness (hereafter referred to by simply 'consciousness') is beyond computation:

4.3 The Argument From Irreversibility

Fix some arbitrary human person; call him 'Bob.' And suppose that Bob is conscious from t_1 to t_{44}. From Proposition $1'$ it follows directly (by elementary first-order logic) that there is a TM M_{Bob}, identical to Bob, such that $C_1 \vdash_{M_{Bob}} C_2 \vdash_{M_{Bob}} \cdots \vdash_{M_{Bob}} C_j$, where this computation is identical to Bob's consciousness from t_1 to t_{44}. But Theorem 1 implies that some algorithm $\mathcal{A}$ reverses the computation in question. Hence, by Leibniz' Law, $\mathcal{A}$ reverses Bob's consciousness from t_1 to t_{44}. But consciousness, whether Bob's or yours or mine, *can't* be reversed. By indirect proof it follows that Proposition $1'$ is false (since Theorem 1, the other possible culprit, is just that: a theorem). And since this proposition is "Strong" AI (or Computationalism) incarnate, it follows that this view is in turn false.

5 Dialectic

5.1 Objection 1

The first objection is rather obvious: "You baldly assert that a stretch of consciousness cannot be reversed. But the irreversibility of consciousness is far from self-evident. I can certainly imagine a person coming into or passing out of consciousness — which seems to make it a reversible process. Do we not also say that people can 'change their minds'?"

The first problem with this objection is that it appeals to senses of 'consciousness' other than the sense we are employing. As we said above, we are specifically considering the question of whether subjective awareness (P-consciousness) over some stretch of time can be reversed. When Jones admits that he has "changed his mind," he isn't saying anything remotely like, "I just experienced the beauty of a sunset backwards." And when we say that Jones received a blow that caused him to lose consciousness, and that presently he was once again alert, we are saying at most that Jones was at one point in such and such a state of P-consciousness, at a subsequent point *not* in a state of P-consciousness, and then thereafter once again in some state of P-consciousness.

As to whether a stretch of P-consciousness is irreversible, we concede *this* much: the proposition in question isn't self-evident *for those who haven't contemplated its truth value*. In order to begin to gain an appreciation for the plausibility of the claim that consciousness is irreversible, one has but to try to pull off the reversal in one's own case. We ask, accordingly, that you indulge in a bit of introspection; we ask that you tackle the challenge of reversing a stretch of your own mental life.[9] In order to fix things a bit, let's suppose that you are quite powerfully affected by hearing Beethoven's Ninth; suppose, more specifically, that you feel energized, deep down, when listening to an initial segment of the choral part of the Ninth during the interval of time t_1 to t_{10}. (If you've never had this experience, merely substitute

[9]It's important to realize that we're talking about your *mental* life: we're not talking about reversing some sequence of physical actions which can be described in such as way as to imply that it is a discrete chain. For example, suppose that you move block a from on top of block b to a position next to block c, and then move block c to on top of block b. You might say that this sequential action is easily reversed by first moving block c to its original position, and then moving block a on top of block b. Though as a matter of fact — as we shall see below — you haven't truly reversed the sequence *qua* physical process, the present point is that we are concerned with reversing a stretch of conscious experience, not a sequence of ostensibly discrete bodily actions.

something similar.) Now, remember your experience during this stretch, fix it in your mind, and proceed to reverse it — live it backwards, so to speak.

What happened? Not much, we wager. Part of the problem seems to be that a stretch of conscious experience can be what might be called *indivisible*. Perhaps if you could divide your Beethoven experience into subchunks corresponding to t_1 to t_3, t_4 to t_6, and so on, you could — thought-experimentally speaking — reassemble them in reverse. But for many, if not most, stretches of P-consciousness such division isn't possible.[10]

We should not be read as proposing divisibility as a quodlibet around which the entire issue revolves. In fact, we're inclined to think that reversing a stretch of P-consciousness is simply *incoherent*. The challenge under consideration seems to us to be analogous to the challenge of imagining that $2 + 2 = 7$, or that *modus ponens* is false. Such things can't be imagined — because they are incoherent (to use a philosophical concept, they are *logically impossible*).

Now to say that a reversed experience of the Ninth is logically impossible is of course to make an exceedingly strong claim. For our main argument it suffices that such a reversed experience be *physically* impossible — a more circumspect claim, and one we as a matter of fact retreat to below. However, after meditating on challenges like the one just posed, and on some states of affairs which are agreed to be logically impossible, it really does seem to us (and to those "subjects" who, at our request, have genuinely attempted to conceive of some reversed stretches of P-consciousness) that the kind of reversal being called for in our challenge *is* logically impossible. We have no outright *proof* that P-consciousness is irreversible in this strong modal sense. Our rationale involves the observation that there are states of affairs deemed by all to be logically impossible (e.g., there being a book both 150 and 200 pages in length at the same time) not because they entail some

[10]Compare this sort of indivisibility with the type Descartes famously ascribed (perhaps incorrectly) to the mind when he said:

> In order to begin this examination, then, I here say, in the first place, that there is a great difference between mind and body, inasmuch as body is by nature always divisible, and the mind is entirely indivisible. For, as a matter of fact, when I consider the mind, that is to say, myself inasmuch as I am only a thinking thing, I cannot distinguish in myself any parts, but apprehend myself to be clearly one and entire; and although the whole mind seems to be united to the whole body, yet if a foot, or an arm, or some other part, is separated from my body, I am aware that nothing has been taken away from my mind. ([13], p. 196)

violation of a truth in logic or mathematics, but rather because when one ponders them, nothing can be envisaged. The same "nothingness" seems to attach to experiencing the Ninth in reverse. And Beethoven's Ninth isn't unique: other more mundane stretches of P-consciousness seem to be just as resistant to reversal: experiences of a dish of fresh strawberries and cream, a vigorous ski run, the reading of a short story, etc. — all these scenarios seem to share the "nothingness" of the 150/200 page book. It is easy enough to imagine swimming the Atlantic unaided, being twenty feet tall, enjoying immortality, finding a counter-example to Boyle's Law, and so on. But it's a tad harder to imagine that the circle has been squared — about as hard, it seems to us, as imagining phenomenal awareness in reverse.

5.2 Objection 2 (The Objection From Physics)

The reply to our rebuttal is likely to be that such phenomenological meanderings are unreliable, and that we seem to be flirting with dualism. People are physical things; consciousness is hence ultimately a physical brain process; ergo, since the Beethoven experience is a stretch of conscious experience, it corresponds to some brain process stretching from t_1 to t_{10} — and surely (so the reply here continues) such a process can be reversed.

Unfortunately, this reply fails. Recall that we explicitly ruled out dualism from the outset by affirming agent materialism. In keeping with this affirmation [an affirmation, specifically, of the proposition we labeled (ii)], and with the spirit of the present objection in mind, let's view the brain in overtly materialistic terms: Suppose that it's composed, not of neurons and dendrites and the like, but of tiny colliding billiard balls. A brain state is thus a "snapshot" of these billiard balls which "freezes" them in a certain configuration. Suppose, further, that we have on hand a microscopic but prodigious helper named Max, a (distant) relative of Maxwell's demon. Max can manipulate neurological billiard balls *par excellence*. It should be possible for Max to reverse billiard ball mentation, if genuine mentation, bound as it is to be infinitely more complex than its homey pool-hall analogue, is reversible. But Max, no matter how clever, no matter how fast, no matter how conscientious, can't pull it off.

In order to see this let's return to Bob's mentation from t_1 to t_{44}. In keeping with our billiards setup, we are now free to view Bob's mentation as the temporally extended interaction of tiny billiard balls inside his cranium. Suppose that Max has kept his eyes on this t_1–t_{44} progression, and that he remembers, impeccably, the passage of these balls through time — their

velocities, spatial and temporal positions, and so on. And suppose that we give Max the opportunity to reverse the mentation. Can he manage? No. And the reason he can't is quite straightforward: Suppose he begins by noting that the 8 ball, which moved from place p_4 to p_6, must be reversed; so he moves the 8 ball from p_6 to p_4 (making sure that the temporal factors match up perfectly with the "forward" recording of the 8 ball). In order to perform his task, the demon must remember what he has done. However, at some point Max's memory will be filled and he will have to erase some stored information. Since erasure is a setting process (i.e., an operation which brings all members of an ensemble to the same observational state) and this operation increases the entropy of the system (billiard balls plus demon) by an amount equal to or greater than the entropy decrease made possible by the newly available memory capacity, the demon cannot execute the reversed mentation.

"Slow down!" exclaims our critic. "If Max's memory is large enough to store the entire sequence of states, then it's large enough to do the reversing; and the obvious way to reverse a sequence of billiard ball collisions is to simply reverse the velocities of the balls in the final state and let the whole system cycle backwards. There is no theorem of statistical mechanics which rules out reversing physical processes."

The central claim here is that since the dynamics of the billiard ball system is completely described by classical statistical mechanics, all we need do to reverse the system is to change the algebraic signs of the motion variables. Any state of the system, so the story goes, is sufficient to allow us to compute any previous and subsequent state.

Unfortunately, this objection conflates the formal system called 'classical' or 'rational mechanics' with a particular instantiation of it by point particles colliding elastically. Reversing is achieved for our critic by manipulating the *equations* of classical statistical mechanics, that is, it is a formal operation, not a physical one. We introduced Max to see what it would take to *actually* reverse a physical process, not simply reverse the formalism used to represent a physical process. Max is supposed to bring into focus the conditions and consequences of executing such a reversal. So it doesn't suffice to say "reverse the velocities of the balls in the final state and let the whole system cycle backwards." We want to know *how* this is done — and this leads us to consider fanciful creatures like Max.

And what Max reminds us of is that while the idealized (point particles, elastic collisions) billiard ball model is logically reversible, the non-ideal (non-point particles, inelastic collisions) is not. The former is logi-

cally equivalent to the formalism it re-represents, and there is a one-to-one correspondence between the discrete states of the idealized model and the continuous state of the mathematical description. The non-ideal model is non-ideal precisely because it lacks those characteristics: there is no simple one-to-one correspondence between an "inelastic collision" and the formalism which tries to capture it. And this is why Max cannot win when working in the real world of friction, dissipation, inelastic collisions, uncertain initial conditions, etc. His memory is soon exhausted, he needs to erase, but the erasure only causes him further difficulties and increases his task.

Our metaphorical Max can be supplanted with a technical discussion making the same point, but such a discussion would quickly render the paper inaccessible to most. This discussion would begin by charting the research devoted to instantiating purely conceptual, reversible models in terms of physical models. This research has lead to a family of devices with intriguing properties; one such device is is the **ballistic computer**, which shows, *in principle*, how a computation can be performed without dissipating the kinetic energy of its signals. The device employs a set of hard spheres and a number of fixed barriers with which the "balls" collide and which cause the "balls" to collide with each other. The collisions are elastic, and between collisions the balls travel in straight lines with constant velocity according to Newton's second law. But what happens when one departs slightly from the idealizations upon which the ballistic computers is based? We find what Max already showed us: when we move from the realm of the ideal and ask how ballistic computers might be physically embodied, we immediately confront two problems: (i) the sensitivity of the ballistic trajectories to small perturbations and (ii) how to make the collisions truly elastic. Initial random errors in position and or velocity of one part in 10^{15} are successively magnified with each generation of collisions, so that after a few dozen generations, the trajectories cannot sustain a computation. Even if we had perfect accuracy and a perfect apparatus, the ballistic computer would still be subject to fluctuating gravitational forces in the environment, so that after a few hundred collisions, the "perfect" trajectories would be spoiled. This dynamical instability of the ballistic computer could be countered by corrections after every few collisions, but then we would no longer have a thermodynamically reversible device: the computer becomes dissipative and logically irreversible. (For a detailed survey of the "Max phenomenon," in the form not only of ballistic computers, but also the enzymatic Turing Machine and the Brownian computer, see [5]. For a detailed, book-length treatment of irreversibility in general, see the book one of us (Zenzen —

with Hollinger) has written on the subject: [42].)

Some skeptics might persist in objecting that our argument founders in its use of physics — and the dialectic could go on, at the cost of producing a paper suitable only for consumption by students and practitioners of physics. Someone might say, for example, that with the possible exception of some rare weak interactions, all of fundamental physics is thought to be reversible in the most straightforward sense. One of the central puzzles of statistical thermodynamics, so the objection goes, is exactly to explain why the world seems to contain irreversible processes when at base it doesn't.

The puzzle here alluded to gives rise to a fundamental rift in physics between those who hold that the world, despite appearances, can't contain irreversible processes (because the formalisms designed to capture the microlevel imply that all physical processes are in principle reversible), and those who hold that no one ought to deny what appears to be clear as day (for reasons of the sort canvassed above in our discussion above of ballistic computers), viz., that *real* physical processes are irreversible. And because there is such a rift, the present objection is uncompelling. In order to see this, first consider someone who, taking P-consciousness seriously, makes a careful case C for property dualism (the best example is probably [22]). Now consider an eliminative materialist who literally denies the existence of P-consciousness (e.g., Dennett [12]). And next consider someone who, apprised of this clash, objects to C as follows. "Look, one of the great puzzles in philosophy is that there appears to be this phenomenon called 'P-consciousness' which resists reduction to purely physical terms. But by the tenets of physicalism, the world contains nothing that cannot be captured in physical terms." This objection leaves C intact for all those who, unlike the eliminative materialist, haven't unalterably placed their bets. Similarly, our argument is intact for all those who haven't permanently affirmed the view that certain formalisms in use in physics imply that nothing is irreversible.[11]

[11] Additional objections are possible, but not powerful. For example, someone might claim that appeals to thermodynamics have no place here since neither Max nor the system being manipulated need be closed. The short answer to this is that only equilibrium thermodynamics requires closed systems. Indeed, our arguments depend on recognizing important differences between equilibrium and non-equilibrium situations. Consciousness necessarily involves departure from equilibrium; computation doesn't.

5.3 Objection 3

At this point those who would resist our argument might turn away from attacks that spring from physics, and claim instead that we are attacking a "straw man:" They might say: "Look, there are plenty of AI and Cog Sci researchers who don't affirm anything like Propositions 1 and 1′. In fact, one can be a darn good AI researcher and not take a stand on the relation between computation and cognition."

This objection is of course easily cast aside — because it fails to take account of the target we have set for ourselves. We are not targeting a brand of AI concerned (say) exclusively with engineering a computational correlate to the olfactory component of rat brains.[12] We are concerned with what, following tradition, we've called "Strong" AI and Cog Sci: a discipline which aims at replicating human cognition in part by identifying, scientifically, cognition with computation.[13] Propositions 1 and 1′ simply reflect our focus on this brand of AI/Cog Sci.

5.4 Objection 4

Against Proposition 1′ (and, for that matter, Proposition 1) it might be said, "This proposition leaves out of the picture something we know to be essential for mentation of the sort we humans enjoy, viz., interchange with the environment. An agent's mentation over some stretch of time doesn't consist solely in computation divorced from the surrounding environment, it consists of computation which reflects a symbiosis with the environment — because human persons have sensors (for taking in information from the environment) and effectors (for manipulating the environment)."

Though we have argued elsewhere that transaction with the "outside" environment is, at least in principle, entirely superfluous when it comes to consciousness,[14] we grant here for the sake of argument the claim that

[12]One of us (Bringsjord) conducts a lot of "Weak" AI, in the form of an attempt to engineer systems capable of autonomously generating stories. See, for example, the programs featured in [2a].

[13]We make veiled reference here to a distinction between *simulation* and *replication*. Details of the distinction aren't necessary for this paper; a rough-and-ready characterization suffices. Accordingly, we say that to simulate some human chess player via AI techniques would be to produce a computer program whose overt behavior (i.e., actual chess moves) resembles that of some human, whereas to replicate Kasparov would be to construct an artifact that literally has the same inner life he has, the same emotions, plans, experiences, mental images, memories, etc.

[14]In [2g] we devise and exploit a variation on the classic brain-in-a-vat gedankenexperi-

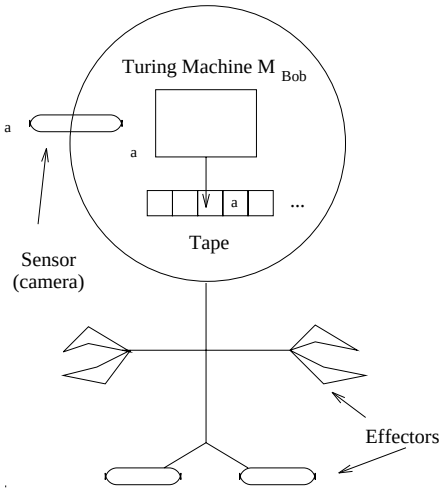


Figure 3: Bob as Transactional Turing Machine

sentience requires such interchange. This concession does imply that Propositions 1 and $1'$ are indeed unacceptable. But (as we indicated in a slightly different context: [2f]), it should be easy enough to remedy the situation. We have only to build in the environmental interchange to Proposition $1'$. In order to do so, let M^*x hold iff x is a TM with sensors and effectors; then we have both a corresponding picture (Figure 3) and a new proposition, viz.,

Proposition 1″. $\forall x\,(Px \wedge x$ is conscious from t_i to $t_{i+k} \Rightarrow \exists y\,(M^*y \wedge x = y \wedge C_j \vdash_y C_{j+1} \vdash_y \cdots \vdash_y C_{j+k}))$, where this computation — partly determined by causal interaction with the environment — is identical to the consciousness x enjoys through $[t_i, t_{i+k}]$.

Now, in order to rebut Objection 3, let's simply expand Max's powers: give him control not only over Bob's "neuro" billiard balls, but over the homey analogue for the molecular motion involved in the use of sensors and effectors. (We leave the details of the more elaborate analogue to your imagination, but a good start is had by simply assuming that the neuro billiard balls behave the way they do in part because billiard balls from the outside interact with them.) Does this change the picture in any way? No. In fact, Max's handiwork is now all the *more* irreversible, for reasons already

ment in order to make the case for this view. In [2c] one of us (Bringsjord) argues, *contra* Harnad [17], that Turing Testing, in order to test for consciousness, needn't include a test for the ability of a would-be AI to interact with its environment.

covered; and so reasoning which parallels our argument in the obvious way arises.

5.5 Objection 5

This objection, more subtle and powerful than its predecessors, begins by bringing before us the argument as it stands given the dialectic to this point:

> Fix some arbitrary human person; call him 'Bob.' And suppose that Bob is conscious from t_1 to t_{44}. From Proposition $1''$ it follows directly (by elementary first-order logic) that there is a sensor-and-effector-outfitted (physical) TM M_{Bob}, identical to Bob, such that $C_1 \vdash_{M_{Bob}} C_2 \vdash_{M_{Bob}} \cdots \vdash_{M_{Bob}} C_m$, where this computation — partly determined by causal interaction with the environment — is identical to Bob's consciousness from t_1 to t_{44}. But Theorem 1 implies that some algorithm $\mathcal{A}$ reverses the computation in question. Hence, by Leibniz' Law, $\mathcal{A}$ reverses Bob's consciousness from t_1 to t_{44}. But consciousness, whether Bob's or yours or mine, *can't* be reversed (as can be seen both via introspection and a look at the elementary physics of irreversibility). By indirect proof it follows that Proposition $1''$ is false (since Theorem 1, the only other possible culprit, is just that: a theorem). And since this proposition is "Strong" AI (or Computationalism) incarnate, it follows that this view is in turn itself false.

Now, the new objection runs as follows: "The inference at *...But Theorem 1 implies that...* is invalid, because Theorem 1 applies to abstract mathematical objects (of a sort studied in *theoretical* computer science), whereas M_{Bob} is now, given the rebuttal to Objection 3, a *physical* object. To link the two as the argument now does, to ascribe to M_{Bob} properties possessed by *mathematical* machines, is to commit a fallacy."

The first problem plaguing this objection is that Strong AI and Cog Sci are founded upon the notion that there is a level of analysis, in the study of the mind, which corresponds to computation. As Dennett [11] recently puts it:

> The central doctrine of cognitive science is that there is a level of analysis, the information-processing level, intermediate between the phenomenological level (the personal level, or the level of consciousness) and the neurophysiological level (p. 195).

It may well be, for example, that at the neuromolecular level, computation is a dispensable notion. It certainly seems that at the phenomenological level computation is at least unwieldy, to say the least. [We routinely give exclusively "folk-psychological" (= phenomenological) explanations of human behavior, as when, e.g., we say such things as that "Bob left the room because he wanted to."] But the information-processing level, the level at which AI and Cog Sci is canonically done, is a different story: it's a level surely governed by results in computability theory; and one such result is none other than our Theorem 1.

There will doubtless be those inclined to sustain the fight for the position that an unbridgeable chasm separates computability theory from M_{Bob} (and the like), so let's explore, in a bit more detail, the relation between such results and the "real" world.

Some well-known theorems in computability theory, in particular those concerning *un*computability, clearly enjoy direct connections with (or at least have immediate implications concerning) the physical world. Consider, for example, the halting problem.

In order to briefly review this problem, we begin by writing

$$M : u \longrightarrow \infty$$

to indicate that TM M goes from input u through a computation that never halts. We write

$$M : u \longrightarrow \textbf{halt}$$

when machine M goes from input u through a computation that *does* halt. And we write

$$M : u \longrightarrow v$$

when machine M goes from input u through a computation that prints out v and then halts.

Let the traditional property of decidability be handled by way of the symbols **Y** ("yes") and **N** ("no"). (For example, the problem of whether a given object a is a member of a set A is decidable iff some machine M' exists which is such that

$$M' : a \longrightarrow \mathbf{Y} \text{ iff } a \in A$$

$$M' : a \longrightarrow \mathbf{N} \text{ iff } a \notin A.)$$

Finally, let n^M represent an encoding, in the form of a natural number, of machine M.[15]

The halting problem now amounts to the theorem (call it 'HP') that it's not the case that there is a TM M such that for every input u, and every M^*:

$$M : n^{M^*} \longrightarrow \mathbf{Y} \text{ iff } M^* : u \longrightarrow \mathbf{halt}$$
$$M : n^{M^*} \longrightarrow \mathbf{N} \text{ iff } M^* : u \longrightarrow \infty$$

Now that we have HP firmly on the table, let's return to our dialectic. The question before us is whether there is an unbridgeable chasm separating HP from the physical world. Put another way: Does HP apply to the physical world, as well as the mathematical?

The answer is an unwavering "Yes." In fact, HP can be "physicalized," in the sense that it can be reworded so as to make a most concrete assertion. One possibility is what we might call HP^P, which says that it's not *physically* possible that one *build* a TM M (or, if you like, a *computer*) such that for every input u (in the "real world" sense of 'input' attached to the word in, say, business and engineering applications), and every *physical* M^*:

$$M : n^{M^*} \longrightarrow \mathbf{Y} \text{ iff } M^* : u \longrightarrow \mathbf{halt}$$
$$M : n^{M^*} \longrightarrow \mathbf{N} \text{ iff } M^* : u \longrightarrow \infty$$

Since HP applies to the corporeal world in the form of HP^P the objection in question looks to be evaporating.

But perhaps it will be said, "You've just been lucky with HP. This theorem does link the 'Platonic' realm with the corporeal one; with this I agree. But this is a coincidence. Other theorems, and in fact perhaps the *bulk* of computability theory, stands inseparably apart from the physical realm. If I had to guess, I would say that HP turned out to apply because it's a *negative* result. Notice that you yourself turned first to *un*computability. Things will turn out differently for theorems that are not negative. And when things do turn out this way, we can return to scrutinize Theorem 1 and its relatives, which are themselves not negative."

Well, let's see if this is right; let's pick a theorem which isn't negative. Specifically, let's consider a fact mentioned above, viz.,

[15] One method for such encoding comes via gödel numbers. For example, see [15].

"Theorem" 2.[16] Turing Machines with multiple tapes are no more powerful than standard one-tape TMs.

For any fixed natural number k, a **k-tape Turing Machine** has k two-way infinite tapes, each of which has its own read/write head. We assume that a k-tape TM can sense in one step the symbols scanned by all heads and, depending upon what those heads find, can proceed with standard actions (erase, write, move right or left).

Though a fully worked out proof of Theorem 2 is perhaps a bit daunting in its detail, the key idea behind the proof is actually rather simple. (Specifying the proof, once one grasps this key idea, is downright tedious.) It is that k tapes can be converted into one tape with k "tracks." An example will make the idea clear: Suppose that we have a 3-tape TM whose configuration, at some moment, is captured by

...	0	0	1 ↑	0	0	0	...

...	0	0	1	1	1 ↑	0	...

...	1 ↑	0	1	0	0	0	...

Then the algorithm at the heart of Theorem 2 would simply convert this into one tape with three tracks (where the single-tape TM's alphabet is suitably composed), as in

...	0	0	1 ↑	0	0	0	...
...	0	0	1	1	1 ↑	0	...
...	1 ↑	0	1	0	0	0	...

Now, the question we face is whether

"Theorem" 2^P. Physical Turing Machines with multiple tapes are no more powerful than standard one-tape physical TMs.

[16] We place scare quotes around 'Theorem' because we describe the result very informally. For a more precise statement of the theorem, as well as a more precise account of the proof than what we provide below, see [24].

is true.[17] And the correct answer would seem to be an affirmative one — because it seems physically possible to follow the algorithm at the heart of the proof of Theorem 2 in order to convert a physical k-tape TM into a single tape physical TM.[18]

Objection 4, then, is beginning to look like a dead end. There does appear to be a link between computability theory and the physical world sufficiently strong to undergird the Argument from Irreversibility. And this is just what we would expect, given that computability theory provides much of the mathematical framework for Computationalism.

At this point some may be inclined to sustain Objection 4 in the following manner: "The two of you say that computation is reversible, and with this I heartily agree. But you also go on to point out that no physical process is reversible; and you then capitalize on this fact. But what you fail to appreciate is that at the information-processing level to which Dennett has drawn our attention consciousness *is* reversible. The problem is that you have surreptitiously moved at the same time to a level *beneath* this level, the level of Max and entropy and thermodynamical equilibrium, which is indeed a level where reversibility fails to apply. Your argument is nothing more than sleight-of-hand."

The problem with this objection is that it conveniently ignores the fact that Computationalism is wed not only to information processing, but also to agent materialism, the view that cognizers are physical things, and that therefore cognition is a physical process. In light of this, introducing at least elementary considerations from physics, as we have done, is not only natural, it's unavoidable. And, as we have shown, once these considerations are introduced, the Argument From Irreversibility is off and running.

The situation can be specified by returning to the adumbration of the argument we offered at the outset of the paper. There, as you may recall,

[17] We leave aside a complexity-based construal of 'power.' Obviously, a physical TM with multiple tapes could sometimes solve problems *faster* than an ordinary one-tape TM. However, there are no problems which are unsolvable by a standard TM yet solvable by a multi-tape machine.

[18] It's easy enough to imagine the details of this conversion in some cases. For example, suppose that we present you with a 4-tape TM roughly in the the form of a model railroad set. That is, suppose that: the tapes are divided into squares; the read/write head, for each tape, is a lone boxcar; there is some way to give simple instructions to this railroad TM; etc. Now imagine converting this physical TM to a one-tape TM via the trick at the heart of Theorem 2. You would have to find some way to link the tracks together into one unit. If you have any experience with model railroading, you probably can visualize yourself tackling the process.

we said that the argument would involve a quartet of propositions numbered (i) through (iv), now unpacked as

(i) Theorem 1;

(ii) Agent Materialism;

(iii) P-consciousness is irreversible, as can be seen by both introspection and an analysis informed by elementary physics; and

(iv) Proposition 1″.

The problem for Computationalism is that the denial of (iv) is entailed by {(i), (ii), (iii)}.[19]

It's important to note that our conclusion needn't worry those who seek to pursue only "Weak" AI, essentially the program satisfied with engineering intelligent systems — systems not intended to replicate those properties (e.g., self-consciousness, command of a language, autonomy, etc.) traditionally thought to be constitutive of personhood. Weak AIniks are free to reject Proposition 1″.[20] We return to this point at the end of the paper.

[19]This is as good a place as any to point out that our argument is of necessity a good deal trickier than this one: "TMs are abstract entities that do not exist in space/time and between which only logical relations obtain, while minds are causal systems in space/time. Logical processes are reversible wile causal processes are not). Therefore minds are not TMs." Our argument is trickier because Computationalism doesn't hold that minds (or persons) are abstract entities. On the contrary, computationalists as a rule hold that minds are physical things. They must also hold, however, that minds, as physical things, abide by the principles of computation — and this is what gets them into trouble, as we are in the process of showing. The simplest distillation of our argument is that it is a proof of inconsistency holding between four propositions, as we indicated at the outset, and as we explain in more detail in the present section.

[20]They are *not* free to reject Theorem 1, however. Two different and determinate levels may be addressed by Theorem 1 (and the like): perhaps there is the purely mathematical level of the theorem itself, and and then perhaps also what might be called the "logic gate level" of computer engineering — a level indispensable for Weak AI. Computer engineers have on hand physical instantiations of Turing Machines which they can combine in order to generate increasingly sophisticated computational artifacts, but there is no guarantee, say, that such engineers will be familiar with the purely formal set theoretic definition of a TM and a TM configuration. Nonetheless, both theoretical computer science *and* computer engineering is constrained and informed by Theorem 1 and its relatives. Work in AI and Cog Sci, whether of the "Weak" or "Strong" variety, is carried out at both of these levels. To abandon these levels and carry out a research program which in its entirety is "below" them, say research exclusively at the neuromolecular level, is to abandon AI for a variant on bio-engineering.

5.6 Objection 6

The next objection marks a return to physics: "You assume that all physical processes are irreversible. But they are not. In fact, under certain conditions, namely those where the system is allowed to move from state A to state B in such fashion as to always be arbitrarily close to thermodynamic equilibrium, it's possible to execute a reversed process. All we have to do is avoid generating or allowing the system to generate an entropy increase which is transferred to the environment and is thus not retrievable. One way to do this is to proceed very slowly so as to not generate heat, in other words, we execute an **adiabatic process**."

This objection flies in the face of what we said above in Objection 2 regarding Max and ballistic computers, which are designed to capitalize on such tricks as adiabatic processes.[21] But let's suppose for the sake of argument that the barriers to real, physical irreversibility described in connection with Max and ballistic computers can somehow be surmounted. Under this charitable supposition, how does the present objection fare? Not well; here's why.

Again, let's suppose for the sake of argument that an automaton can be made reversible at every step *in the real physical world*, and that this allows not only an *in-principle* thermodynamically reversible automaton, but a "real world" one: a concrete one that saves all intermediate results, avoids irreversible erasure, prints the desired output and reversibly disposes of all undesired intermediate results as it retraces the machine's configurations in reverse order. To accomplish this, in keeping with the ideas behind ballistic computers, the automaton must be kept or keep itself arbitrarily close to thermodynamic equilibrium. For erasure to be thermodynamically reversible, the initial ensemble of memories must be distributed equally among possible states of phase space. Let's call such a process **r.w.i**, for "real-world irreversible." . Now, given that consciousness isn't currently associated with r.w.i. processes, why isn't the objection a complete non-starter? The answer must be that somehow — at least in the realm of thought and thought-experiment — the processes associated with consciousness can, at least in principle, become r.w.i. But notice that since the objective of our opponent is to show that reversing *consciousness* is doable, if the present objection is to have a chance of succeeding, it must assert not only that the neurological processes in question can become r.w.i, but also that they could become

[21] Again, for a detailed discussion of Max, ballistic computers, and other, similar devices, see [6] and [4].

r.w.i. *without thereby preventing the rise of the consciousness with which they are connected*. And so a natural question arises: Is there any reason to think that the speed at which neurological processes unfold can be inseparably bound up with the consciousness they underlie? Indeed there is, as we now proceed to explain.

Suppose that Sandra enjoys a stretch of consciousness from t_1 to t_4 which centers around an appreciation of her own brain activity during this time. (In order to make it easy to envisage the situation, imagine that Sandra's doctors are concerned that she might have a brain tumor, and hence submit her to various brain scanning technologies.) Suppose, in addition, that she not only observes her brain activity during this time, but that she apprehends the speed at which these neurological processes proceed. (To make the scenario vivid, you might think of her watching a PET scan as a digital readout ticks off the scan's duration.) It's impossible, *ex hypothesi*, that Sandra's brain processes through $[t_1, t_4]$ be "r.w.i.-slowed" without destroying the consciousness with which they are associated in the "normal speed" mode. The problem, specifically, is that if 'Bob' is replaced with 'Sandra' in the Argument From Irreversibility (see the "double-boxed" version of the argument given above), the present objection is impotent.[22]

Generally speaking, could a mind/brain ever meet r.w.i. conditions or their equivalents? In particular, could the mind/brain be kept at or near equilibrium so that a mentation could be reversed? Given that the firing of a single neuron is decidedly *not* an adiabatic process, and given that a mind/brain at or near equilibrium is a *dead* brain (or, at best, a "blank" mind), we are compelled to conclude that cognition cannot be reversed because *the initial conditions* required for reversal cannot be met. It seems *very* unlikely that the physical processes of the central nervous system which sustain P-consciousness can be adequately modeled even by the most sophisticated ballistic computer (Brownian computer, etc.).

In sum, the difference between computation and cognition is this: Computation is logically reversible à la Theorem 1. And computation can

[22]This is as good a place as any to register our prediction that at this juncture those desperate to dodge our argument might say about the Sandra thought-experiment what those who defend the coherence of time travel have said about the so-called "Grandfather Paradox" (GP, in a word, being that if time travel if possible, then you could go back in time and kill your grandfather, but then how could you be around to do the killing?), namely that the Sandra case just won't ever happen. This move is laughably *ad hoc*. Besides, Sandra's P-consciousness isn't really any different than garden variety stretches of P-consciousness evoked by watching a movie played from a VCR whose digital clock keeps normal time.

be modeled by devices (e.g., the ballistic computer) which approximate r.w.i. processes. Cognition, however, including as it does P-consciousness, is phenomenologically, experientially, and conceptually *irr*eversible, and *cannot* be modeled by devices which approximate physical reversibility.

5.7 Objection 7

The next objection we consider marks an appeal to connectionism; it runs as follows. "Your argument, I concede, destroys Computationalism. But that suits me just fine: I don't affirm this doctrine, not in the least. For you see, cognition isn't computation; cognition, rather, is suitably organized processing in an architecture which is at least to some degree genuinely neural. And I'm sure you will agree that a Turing Machine is far from neural! Neural nets, however, as their name suggests, *are* brain-like; and it's on the shoulders of neural nets that the spirit of 'Strong' AI and Cognitive Science continues to live, and indeed to thrive."

This objection entails a rejection of Proposition $1''$. As such, it succeeds in dodging the Argument From Irreversibility — since this argument has this proposition as a premise. However, the argument can be rather easily revived, by putting in place of Proposition $1''$ a counterpart based on neural nets, viz. (where N^*x iff x is an corporeal neural net),

Proposition 2. $\forall x\,(Px \wedge$ x is conscious from t_i to $t_{i+k} \Rightarrow \exists y\,(N^*y \wedge x = y \wedge y$ passes through some process — partly determined by causal interaction with the environment — which is identical to the consciousness x enjoys through $[t_i, t_{i+k}])$.

This proposition is affirmed by connectionists (of the "Strong" variety, anyway).[23] But Proposition 2, combined with a chain of equivalence holding between neural nets, cellular automata, k-tape Turing Machines, and standard TMs, is enough resurrect the Argument from Irreversibility in full force. The chain of equivalence has been discussed in detail by one of us (Bringsjord) elsewhere [2f], in a paper which purports to show that connectionism, at bottom, is orthodox Computationalism in disguise.[24] Here, it will suffice to give an intuitive recapitulation of the chain in question —

[23] One of us (Bringsjord) distinguishes between various sorts of computationalists and connectionists in [2f].

[24] This paper is based on the classic statement of connectionism given by Paul Smolensky [38], [39].

a chain which establishes the interchangeability of neural nets and Turing Machines.

Before we sketch this chain, let's pause to make clear that it underlies a certain proposition which can be conjoined with Proposition 2 in order to adapt the Argument From Irreversibility. This proposition is

Proposition 3. $\forall x\,((N^{*}x \ \wedge \ x$ passes through some process through $[t_i, t_{i+k}]) \ \Rightarrow \ \exists y\,(My \ \wedge \ C_j \vdash_y C_{j+1} \vdash_y \cdots \vdash_y C_{j+p})$, where this computation is identical to the process x enjoys through $[t_i, t_{i+k}])$.

It's easy to prove in elementary logic (using such rules as universal elimination and *modus ponens*) that Proposition 2, conjoined with Proposition 3, reenergizes the Argument From Irreversibility. But why is Proposition 3 true? In order to answer this question, we need to look first at neural nets. After that, even a casual look at "two-dimensional" TMs should make it plain that Proposition 3 is true.

Neural nets are composed of **units** or **nodes**, which are connected by **links**, each of which has a numeric **weight**. It is usually assumed that some of the units work in symbiosis with 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:

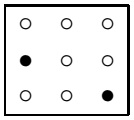
$$a_i = g(in_i) = g\left(\sum_j W_{ji} aj\right).$$

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.[25] (This is supposed to look

[25] McCulloch and Pitts [27] showed long ago that such a simple activation function allows for the representation of the basic Boolean functions of AND, OR and NOT.

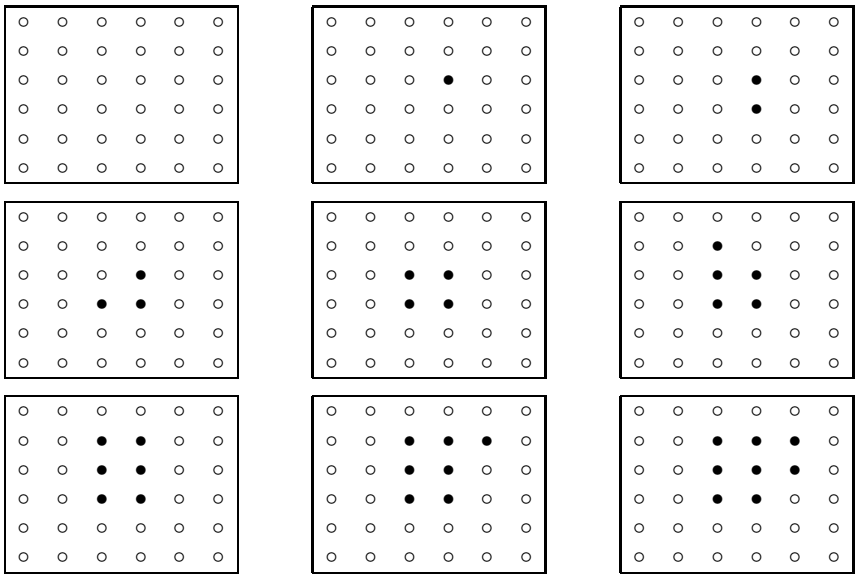
"brain-like" to some degree, given the metaphor that 1 represents the firing of a pulse from a neuron through an axon, and 0 represents no firing.)

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. But no matter what neural net you care to talk about, Proposition 3's deployment of its universal quantifier remains justified. Proving this would require a paper rather more substantial than the present one, but there is a way to make the point in short order. The first step is to note that neural nets can be viewed as a series of snapshots capturing the state of its nodes. For example, if we assume for simplicity that we have a 3-layer net (one input layer, one "hidden" layer, and one output layer) whose nodes, at any given time, or either "on" (filled circle) or "off" (blank circle), then here is such a snapshot:



As the units in this net compute and the net moves through time, snapshots will capture different patterns. But Turing Machines can accomplish the very same thing. In order to show this, we ask that you briefly consider **two-dimensional** TMs. We saw k-tape TMs above; and we noted the equivalence between these machines and standard TMs. One-head two-dimensional TMs are simpler than k-tape machines, but (in the present multidisciplinary context) more appropriate than their standard k-tape cousins. Two-dimensional TMs have an infinite two-dimensional grid instead of a one-dimensional tape. As an example, consider a two-dimensional TM which produces an infinite "swirling" pattern. We present a series of snapshots (starting with the initial configuration, in which all squares are blank, and moving through the next eight configurations produced by the "swirl" program) of this machine in action.[26] Here's the series:

[26]The machine itself, as a series of quadruples (with 0 for blank, 1 for filled, R and L for "right" and "left," resp., A for "any," etc.) is (read left to write, top to bottom):



The point of this series of snapshots is to convey that snapshots of a neural net in action can be captured by a one-head two-dimensional TM (and, more easily, by a k-tape, k-head machine). Hopefully you can see why this is so, even in the absence of the proof itself. The trick, of course, is to first view the neural net in question as an $n \times n$ array (we ignore the part of the infinite grid beyond this finite array), as we did above. Of course, it's necessary to provide a suitably enlarged alphabet for our neural-net-copying TM: it will need to have an alphabet which contains a character corresponding to all the states a node can be in. For this reason, our swirl TM is a bit limited, since it has but a binary alphabet. But it's easy enough to boost the alphabet (and thereby produce some entrancing pyrotechnics).[27]

The proponent of Objection 7 might say that a sequence of configurations of a Turing Machine M proposed to capture a neural net N as it passes through time isn't complete, because such a sequence includes, at best,

1011	1102	2012	21L3	3013
31U4	4014	41U5	5015	51R6
6016	61D7	70U9	71U8	8AR6
9AD10	100110	101L11	111R9	110R12
12AL13	130113	131U14	141D12	140D15
15A16	160116	161R17	171L15	170L8

The implementation of this machine — in Turing's World^{TM} — sent upon request.

[27]Readers wanting to see some of them are encouraged to consult Poundstone's [30] classic discussion of Conway's [7] Game of Life.

only *discrete* "snapshots" of the *continuous* entity N. But suppose that the sequence includes snapshots of N at t $(= N_t)$ and $t+1$ $(= N_{t+1})$, and that our opponent is concerned with what is left out here, i.e., with the states of N between t and $t+1$. The concern is easily handled: one can make the interval of time during which the state of N is ignored arbitrarily small. Doing so makes it the case that the states of N which had formerly been left out are now captured.

5.8 Objection 8

The next objection is based on an analogy: "Look, a chess game is nothing but a sequence of moves of the pieces in accord with the rules of chess. No one would claim that the ontological status of chess games is particularly mysterious. The sequence of moves in any chess game is reversible: there exists a sequence which consists in exactly the same positions on the board but in opposite order. The backward sequence is not, of course, a chess game. The 'initial' position is not the legal initial position for chess, and some pieces, e.g., pawns, will move in a way not legally permissible. So the chess game cannot be reversed, in the following sense: the reverse sequence of positions is not a legal chess game. Does one want to use Leibniz's Law to conclude that a particular chess game is not identical to the sequence of moves in it, because the one is reversible and the other isn't? I wouldn't want to conclude this, but in any case it doesn't matter. In whatever sense one cares about it, the chess game is nothing over and above the sequence of moves: if similarly consciousness were nothing over and above a computation, then the defenders of Strong AI win."

The analogy offered here is clever, and if we were to accept it and its implied purely syntactical view of chess, then our master argument would probably be threatened. But if we remember that chess is a sequence of moves made amid rules, intentions, beliefs, plans, goals, and so on, there is no reason to accept an analogy which *ab initio* commits us to a reductionist view of consciousness. In fact, the objection, upon reflection, can be exposed as fundamentally flawed — because it's easy enough to *show*, courtesy of a thought-experiment, that what the discussant takes to be absurd (a chess game is something over and above a sequence of board configurations) is *true*: Begin by fixing some legal sequence S of chess moves from the required initial situation to one in which the white king is checkmated. Suppose that this sequence is one which has never been actualized in the past. Now imagine a chess board lying on the surface of an uninhabited planet, with chess pieces

beside it.[28] Next imagine that, as a wild coincidence would have it, the lifeless winds blowing over the surface of this planet happen to jumble the chess pieces over an interval of time in such a way that S is actualized. Has a chess game taken place on this planet? It would seem not. (If a game took place, who won and who lost?) But then the key premise in Objection 8 (viz., a chess game is nothing over and above a sequence of moves) is in grave doubt, and hence Objection 7 evaporates.[29]

5.9 Objection 9

Here's the next objection: "It's not clear that the sense in which you claim computation is reversible is relevant. You identify a computation as a sequence of total states of a machine, where the total state specifies the machine state, read/write head location, and tape register contents. You then assert that there is *some other* program M' which will cause the computer to run through those states in reverse order. But the *identity* of machine states is not a fact which is separable from the program which is actually running on a machine. Machine state "1," for example, is not that very state because it has a "1" pasted to it, but because of the transitions between it and other states which are caused by the program which the machine is running. So in switching from the original program M to M', we no longer have the same machine states available, so the sequence induced by M' isn't the reverse of the original. And one can't reverse a computation by reversing the program in any interesting way. There is, for example, a simple two-state program which will erase a sequence of 1s of any length, but one cannot 'reverse' it to get a two-state program which will write out such a sequence."

This objection saddles Computationalism with **machine state functionalism** (MSF), according to which our mental states are to be identified with the *machine states* of a TM rather than the *configurations* of such a machine. Unfortunately, while it's true that the output of the algorithm $\mathcal{A}$ of Theorem 1 is never a TM which in any sense "reverses" the machine states of its input, machine state functionalism has long ago been buried; no contemporary computationalist advances this view. (The *locus classicus*

[28] Don't worry about how the chess board and pieces got there in the first place. Perhaps our props were launched into space centuries ago, when the human race still existed. Perhaps the board and the pieces were formed from mud by the random winds invoked below...

[29] It is at any rate undeniable, given this gedankenexperiment, that the ontological status of chess games *is* a bit mysterious.

of MSF is due to Putnam [31], who has himself rejected the doctrine.) The reasons MSF is a carcass are myriad; they are nicely catalogued in Chapter 8 of [41]. One problem with MSF is the apparent unboundedness of human mental states. It has seemed to many that humans can enter any of an infinite number of mental states. (One could believe that 1 is the successor of 0, that 2 is the successor of 1, that 3 is the successor of 2, ..., and so on *ad infinitum*. And of course we would need to consider states involving not only beliefs about arithmetic, but also hopes, fears, dreams, mental images, and so on.) But every TM has a fixed and finite set of machine states (while on the other hand even tiny TMs are capable of entering an infinite number of configurations.) Another agreed upon defect plaguing MSF is that according to it two TMs which compute the same function f but which differ in their machine states and the arcs connecting them (to use the critic's scheme) are classified as giving rise to different cognition. But this implies that if you share with us (say) a love of climbing roses of the "Blaze" color, underlying our attitude must be one TM with the same exact states — which hasn't seemed too plausible to most.

Finally, the present objection is problematic for another reason having nothing to do with the history of Computationalism: Computationalism is the view that cognition (including consciousness) is *computation*, but computation is *not* a machine state (or a collection of such states, or a collection of such states linked by arcs). Computation, in the terms our critic prefers, isn't a program; rather, computation is a program *in progress*. That is, computation is a sequence of configurations [9], [2d], as we have explained above.

6 Uncomputable Cognition

Our principal concern to this point has been to articulate and defend the Argument From Irreversibility. This argument, with respect to Computationalism (whether of the logicist or connectionist variety), is a negative one: its conclusion is that Computationalism is simply false. Do we have anything *constructive* to say? What view of the mind would we put in place of Computationalism? What are the consequences of what we've uncovered for AI and Cog Sci practitioners? Though such questions in many ways take us outside our expertise, we'll venture a brief answer, in the form of a three-point homily.

First, it's important to realize that we consider the view advocated herein

to be the prolegomenon to a more sophisticated science of the mind. Just because cognition is (at least in part) uncomputable doesn't mean that it will resist scientific analysis. After all, computer science includes an entire subfield devoted to the rigorous study of uncomputability. We know that there are grades of uncomputability, we know much about the relationships between these grades, we know how uncomputability relates to computability, and so on; uncomputability theory, mathematically speaking, is no different than any other mature branch of classical mathematics. So, why can't uncomputability theory be linked to work devoted to the scientific analysis of the uncomputable side of human cognition? That it can be so linked is the view developed and championed in a forthcoming monograph of ours [2b].

The second point, which is related to the first, is this: One of the interesting properties of AI research is that people have come to expect that it invariably have *two* sides, a scientific side, and an implementation side. The basic idea is that the scientific side, which can be quite theoretical, ought to be translated into working computer programs. We think this idea, as a general policy, is wrongheaded. Why is it that the physicist can be profitably occupied with theories that can't be implemented, while the AI researcher labors under the onerous expectation that those aspects of the mind which admit of scientific explanation must also admit of replication in the form of computer programs? In short, perhaps we can come to understand cognition and consciousness scientifically (and this would entail, in the present context, exploiting information-processing systems which *aren't* reversible, e.g., "machines" more powerful than Turing Machines), while at the same time acknowledging that we can't build conscious computers.

The third and final point in our sermon is this. Suppose that we're correct; suppose that human cognition is uncomputable, and that therefore it is something no computer can possess. From this is doesn't follow that no system can *appear* to enjoy consciousness. There are some well-known uncomputable functions which many are doing their best to "solve." (One such line of research is the attack on the uncomputable **busy beaver function** [25].) AI, as far as we can see, has never settled the fundamental clash between those who, like Turing, aim only at engineering a device whose behavior is indistinguishable from ours ("Weak" AI), and those who seek not only to create the behavior but also the underlying conscious states which we humans enjoy ("Strong" AI). Nothing we have said herein precludes success in the attempt to engineer a computational system which *appears* to to be genuinely conscious. Indeed, an approach which cheerfully resigns it-

self to engineering behavior only seems to us to be a route worth taking.[30] What we purport to have shown, or at least made plausible, is that the road down which aspiring "person builders" are walking is ultimately a dead end, because no mere computational system can *in fact* be conscious, and this is true in part because while computation is reversible, consciousness isn't.

References

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

[2] Barr, A. (1983) "Artificial Intelligence: Cognition as Computation," in Fritz Machlup, ed., *The Study of Information: Interdisciplinary Messages* (New York, NY: Wiley-Interscience), pp. 237-262.

[3] Barwise, J. & Etchemendy, J. (1993) *Turing's World 3.0* (Stanford, CA: CSLI Publications).

[4] Bennett, C.H. (1984) "Thermodynamically Reversible Computation," *Phys. Rev. Lett.* **53**: 1202.

[5] Bennett, C.H. (1982) "The Thermodynamics of Computation — A Review," *International Journal of Theoretical Physics* **21**: 905-940.

[6] Bennett, C. H. (1973) "Logical Reversibility of Computation," *IBM Journal of Research Development* November: 525-532.

[7] Berlekamp, E., Conway, J., Guy, R. (1982) *Winning Ways*, vol. 2 (NY, NY: Academic Press). See chapter 25 for Conway's description of Life.

[8] Block, N. (1995) "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences* **18**: 227-247.

[9] Boolos, G.S. & Jeffrey, R.C. (1980) *Computability and Logic* (Cambridge, UK: Cambridge University Press).

[2a] Bringsjord, S. & Ferrucci, D. (forthcoming) *Artificial Intelligence, Literary Creativity, and Story Generation: The State of the Art* (Hillsdale, NJ: Lawrence Erlbaum).

[30] It is the route one of us (Bringsjord) follows in the attempt to engineer systems which *appear* to be creative; see [2a].

[2b] Bringsjord, S. & Zenzen, M. (forthcoming) *In Defense of Uncomputable Cognition* (Dordrecht, The Netherlands: Kluwer).

[2c] Bringsjord, S. (1995) "Could, How Could We Tell If, and Why Should — Androids Have Inner Lives," in Ford, K. & Glymour, C., eds., *Android Epistemology*, (Cambridge, MA: MIT Press), pp. 93-122.

[2d] Bringsjord, S. (1994) "Computation, Among Other Things, Is Beneath Us," *Minds & Machines* **4.4**: 469-488.

[2e] Bringsjord, S. (1992) *What Robot's Can and Can't Be* (Dordrecht, The Netherlands: Kluwer).

[2f] Bringsjord, S. (1991) "Is the Connectionist-Logicist Clash One of AI's Wonderful Red Herrings?" *Journal of Experimental and Artificial Intelligence* **3.4**: 319-349.

[2g] Bringsjord, S & Zenzen, M. (1991) "In Defense of Hyper-Logicist AI," *IJCAI '91*, (Mountain View, CA: Morgan Kaufmann), pp. 1066-1072.

[10] Davis, M. and E. Weyuker (1983) *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science* (New York, NY: Academic Press).

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

[12] Dennett, D. (1991) *Consciousness Explained* (Boston, MA: Little, Brown).

[13] Descartes, R. (1911–first edition) *The Philosophical Works of Descartes*, Vol. I, translated by Haldane, E.S. & Ross, G.R.T. (Cambridge, UK: Cambridge University Press).

[14] Dietrich, E. (1990) "Computationalism," *Social Epistemology* **4.2**: 135-154.

[15] Ebbinghaus, H.D., Flum, J., Thomas, W. (1984) *Mathematical Logic* (New York, NY: Springer-Verlag).

[16] Fetzer, J. (1994) "Mental Algorithms: Are Minds Computational Systems?" *Pragmatics and Cognition* **2.1**: 1-29.

[17] Harnad, S. (1991) "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem," *Minds and Machines* **1.1**: 43-54.

[18] Haugeland, J. (1981) *Artificial Intelligence: The Very Idea* (Cambridge, MA: MIT Press).

[19] Hobbes, T. (1839) De Corpore, chap. 1, in *English Works*, ed. Molesworth, reprinted in (1962) *Body, Man and Citizen* (New York, NY: Collier).

[20] 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, NY: Bantam), pp. 631-665.

[21] Hopcroft, J.E. & Ullman, J.D. (1979) *Introduction to Automata Theory, Languages and Computation* (Reading, MA: Addison-Wesley).

[22] Jacquette, D. (1994) *Philosophy of Mind* (Englewood Cliffs, NJ: Prentice-Hall).

[23] Johnson-Laird, P. (1988) *The Computer and the Mind* (Cambridge, MA: Harvard University Press).

[24] Lewis, H. and C. Papadimitriou (1981) *Elements of the Theory of Computation* (Englewood Cliffs, NJ: Prentice-Hall).

[25] Marxen, H. & Buntrock, J. (1990) "Attacking the Busy Beaver 5," *Bulletin of the European Association for Theoretical Computer Science* **40**: 247-251.

[26] Maudlin, T. (1989) "Computation and Consciousness," *Journal of Philosophy* **84**: 407-432.

[27] McCulloch, W.S. & Pitts, W. (1943) "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* **5**: 115-137.

[28] Newell, A. (1980) "Physical Symbol Systems," *Cognitive Science* **4**: 135-183.

[29] Partee, B., Meulen, A. & Wall, R. (1990) *Mathematical Methods in Linguistics* (Dordrecht, The Netherlands: Kluwer Academic Publishers).

[30] Poundstone, W. (1985) *The Recursive Universe* (NY, NY: William Morrow).

[31] Putnam, H. (1960) "Minds and Machines," in his (1975) *Mind, Language, and Reality: Philosophical Papers* vol. 2 (Cambridge, UK: Cambridge University Press).

[32] Rado, T. (1963) "On Non-Computable Functions," *Bell System Technical Journal* **41**: 877-884.

[33] Rogers, H. (1967) *Theory of Recursive Functions and Effective Computability* (New York, NY: McGraw-Hill).

[34] Schacter, D.L. (1989) "On the Relation Between Memory and Consciousness: Dissociable Interactions and Conscious Experience," in *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*, ed. H. Roediger & F. Craik (Hillsdale, NJ: Erlbaum).

[35] Searle, J. (1980) "Minds, Brains and Programs," *Behavioral and Brain Sciences* **3**: 417-424.

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

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

[38] Smolensky, P. (1988) "On the Proper Treatment of Connectionism," *Behavioral & Brain Sciences* **11**: 1-22.

[39] Smolensky, P. (1988) "Putting Together Connectionism — Again," *Behavioral & Brain Sciences* **11**: 59-70.

[40] Soare, R. (1980) *Recursively Enumerable Sets and Degrees* (New York, NY: Springer-Verlag).

[41] Stillings, N.A., Weisler, S.E., Chase, C.H., Feinstein, M.H., Garfield, J.L., Rissland, E.L. (1995) *Cognitive Science: An Introduction* (Cambridge, MA: MIT Press).

[42] Zenzen, M. and Hollinger, H. (1985) *The Nature of Irreversibility* (Dordrecht, The Netherlands: D. Reidel).