

The Modalized Gödelian Argument Against Computationalism*

Selmer Bringsjord

Dept. of Philosophy, Psychology & Cognitive Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu

Kostas Arkoudas

AI Lab
Massachusetts Institute of Technology (MIT)
NE43-803
Cambridge MA 02139 USA
koud@ai.mit.edu

May 25, 2001

*We're indebted to Leslie Burkholder, David Chalmers, Jack Copeland, Martin Davis, Jim Fahey, Ken Ford, Pat Hayes, Andrew Irvine, Bob McNaughton, Marvin Minsky, Kelsey Rinella, Stuart Shapiro, Chris Welty, and Michael Zenzen. Bringsjord would like to express special thanks to Roger Penrose for extemporaneous debate and conversation free of written text of the sort that, by our lights, for reasons revealed below, has hampered Penrose's Gödelian case.

1 Introduction

Four decades ago, J.R. Lucas (1964) expressed supreme confidence that Gödel’s first incompleteness theorem (= Gödel I) entails the falsity of computationalism, the view that human persons are computing machines (e.g., Turing machines). Though his own arguments have not proved to be compelling, Lucas has had an indefatigable defender: Roger Penrose. Penrose’s first attempt to vindicate Lucas was a Gödelian attack on computationalism articulated in *The Emperor’s New Mind* (1989). Unfortunately, even Penrose has admitted that this first attempt fell short — and so he went back to the drawing board, and soon thereafter came charging into the arena again, armed with a more elaborate and more fastidious Gödelian case, expressed in Chapters 2 and 3 of his *Shadows of the Mind* (1994). Yet once again there was failure, as two sustained, painstakingly formal refutations apparently reveal (LaForte, Hayes and Ford 1998, Bringsjord and Xiao 2000). The upshot is that as we pass into the new millennium, Lucas’ confidence is looking more and more like an emotional quirk, and less and less like an attitude born of careful ratiocination.

In this paper we undertake to reverse the situation. In section 2 we encapsulate Penrose’s fundamental errors, by focussing on his most recent version of his Gödelian case. In section 3 we explain that Penrose’s central intuitions about human mathematical reasoning are nonetheless formidable, if interpreted as intuitions about the more-than-mechanical power of human *infinitary* reasoning. In section 4 we present two general avenues for moving from such intuition to a precise Gödelian argument. We take the second of these avenues, which leads to a victorious *modalized* Gödelian refutation of computationalism consistent with Penrose’s intuitions, and that has its roots in “Chapter VII: Gödel” of *What Robots Can and Can’t Be* (Bringsjord 1992), wherein the kernel of a modal argument against computationalism is presented.¹ In the next section of the paper, 5, we defend the key premise in the argument of section 4. In section 6, we consider and rebut some objections to our modal argument. The paper ends with a brief concluding section.

2 A Fatal Problem Infecting Penrose’s Project

This is not the place to revisit all the technical Penrosean errors exposed in the likes of (LaForte et al. 1998, Bringsjord and Xiao 2000, Feferman 1995); such a visit would take all the space we have, and these critiques stand on their own. Fortunately, there is a shortcut we can take here in order to convey the stew Penrose continues to find himself in: Writing in response to critics (e.g., the philosopher David Chalmers, the logician Solomon Feferman, and the computer scientist Drew McDermott) of his *Shadows of the Mind* in the electronic journal *Psyche*, Penrose has offered a Gödelian case designed to improve on the version presented in

¹The present paper makes good, specifically, on the following promissory note:

I should point out that my [just-given] Gödelian argument against the proposition that persons are automata can apparently be modalized . . . In this modal argument the central claim would be only that Ralf is such that it’s *logically possible* that he act as I have described him acting above via the Fixed Point Theorem, whereas no Turing machine is such that it is logically possible that it act in this way. (Bringsjord 1992, p. 264)

SOTM.² Indeed, in this response Penrose gives what he takes to be the perfected version of the core Gödelian case given in *SOTM*. Here is this version, verbatim:

We try to suppose that the totality of methods of (unassailable) mathematical reasoning that are in principle humanly accessible can be encapsulated in some (not necessarily computational) sound formal system F . A human mathematician, if presented with F , could argue as follows (bearing in mind that the phrase “I am F ” is merely a shorthand for “ F encapsulates all the humanly accessible methods of mathematical proof”):

(A) “Though I don’t know that I necessarily am F , I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion “I am F .” I perceive that it follows from the assumption that I am F that the Gödel statement $G(F')$ would have to be true and, furthermore, that it would not be a consequence of F' . But I have just perceived that “If I happened to be F , then $G(F')$ would have to be true,” and perceptions of this nature would be precisely what F' is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F' , I deduce that I cannot be F after all. Moreover, this applies to any other (Gödelizable) system, in place of F .” (Penrose 1996, ¶ 3.2)

Unfortunately, (A) is a bad argument, as is easily seen. In order to see this, let’s follow Penrose directly and set

$$\psi = \text{“}F \text{ encapsulates all the humanly accessible methods of mathematical proof”}$$

and

$$F' = F \cup \{\psi\}$$

What the hypothetical human mathematician can now conclude, on the strength, as Penrose tells us, of Gödel’s work, is that on the assumption that ψ ,

- (1) $G(F')$ is true.
- (2) $F' \not\vdash G(F')$ and $F' \not\vdash \neg G(F')$

The idea is really quite simple. It is that there is a contradiction arising from the fact that the hypothetical mathematician, i.e. F , can conclude that (1) $G(F')$ is true on the one hand, and yet (2), which “says” that F cannot conclude $G(F')$, is true on the other. But wait a minute; look closer here. Where is the contradiction, exactly? There is no contradiction. The reason is that (1) is a *meta*-mathematical assertion; it’s a claim about *satisfaction*. More precisely, where \mathcal{I} is an interpretation of the relevant type, (1) is just

- (1') $\mathcal{I} \models G(F')$ is true.

And for all we know, F can prove (1') while being bound by (2)! So we see here the classic error originating with Lucas himself: Penrose conflates proofs within a fixed system with meta-proofs.³

²The dialectic appeared in 1996 in volume 2 of *Psyche*, which can be accessed via

- <http://psyche.cs.monash.edu>

And of course *Psyche* can be located using any standard search engine.

³As explained in detail in “Chapter VII: Gödel” of (Bringsjord 1992).

3 Taking Penrose’s Intuitions Seriously

Despite succumbing to errors like these, it seems to us that Penrose’s intuitions about the hypercomputational⁴ nature of human mathematical intuition haven’t been fully appreciated. Could it be that Penrose has apprehended the hypercomputational nature of human mathematical reasoning, but is unable to convince others in a rigorous third-person scheme (logic and technical philosophy) that this is the case? That Penrose’s attempts to articulate a Gödelian case against computationalism are fatally flawed, while his intuitions are sound, is a view that Turing himself would have found palatable, at least in general. In his dissertation at Princeton University (Turing 1938) (later published as Turing 1939), Turing distinguished between “intuition” and “ingenuity” in logic and mathematics. He wrote:

Mathematical reasoning may be regarded rather schematically as the exercise of a combination of two faculties, which we may call *intuition* and *ingenuity*. The activity of the intuition consists in making spontaneous judgments which are not the result of conscious trains of reasoning . . . The exercise of ingenuity in mathematics consists in aiding the intuition through suitable arrangements of propositions, and perhaps geometrical figures or drawings. (Turing 1939, p. 214–215)

We believe that while Penrose’s ingenuity, so preeminently crisp in mathematical physics, has failed him in the Gödelian sphere, his intuition in this sphere *is* in fact exactly right. In order show that this position is more than charity, we begin by turning to some of Penrose’s examples of mathematical thinking.

The first example of such thinking given in *SOTM* is in section 1.19, and involves a visualization-based grasping of the arithmetical fact that

$$a \times b = b \times a, \text{ where } a, b \in \{0, 1, 2, \dots\}.$$

As Penrose explains, in the particular case where $a = 3$ and $b = 5$ we have, for $a \times b$:

$$(\bullet \bullet \bullet \bullet \bullet)(\bullet \bullet \bullet \bullet \bullet)(\bullet \bullet \bullet \bullet \bullet)$$

whereas for $b \times a$ we have

$$(\bullet \bullet \bullet)(\bullet \bullet \bullet)(\bullet \bullet \bullet)(\bullet \bullet \bullet)(\bullet \bullet \bullet).$$

But he reminds us that the fact in question can be grasped by visualizing the array

$$\begin{array}{cccc} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{array}$$

and “rotating the image through a right angle in one’s mind’s eye, to see that the array representing 5×3 has the same number of elements as that representing 3×5 ” (Penrose 1994, p. 55). Now we know that it seems to most that this kind of operation is easy enough for computers to carry out. The reason is that computers can process what one of us (1996) has

⁴The term ‘hypercomputation’ is becoming the standard term to denote information processing beyond what a Turing machine can muster. We explicate the term, to some degree, below.

called ‘simple diagrams,’ or S-D’s for short. An S-D is a diagram that can be fully represented in standard first-order logic (FOL); and standard FOL is comfortably processed by standard computers (e.g., see Russell and Norvig 1994). Obviously, arrays can be represented via tuples in FOL. Suppose that $\#$ is a function mapping $n \times m$ arrays of dots into the natural number corresponding to the number of dots in the array. Let the array above be denoted by the constant a . Then $\#(a) = 15$. Next, suppose that function r^{90c} captures the rotation of an array “through a right angle.” Hence

$$\#(a) = \#(r^{90c}(a)).$$

It is also easy enough to give a procedural representation of the objects and operations Penrose describes. For example, a k -tape Turing machine (with multiple tapes) can be programmed to carry out a direct analogue for the mental rotation in question.⁵ All of this isn’t likely to worry Penrose, for it’s the *general* truths that are at issue for him — truths like (where we are assumed to be quantifying over arrays)

$$\forall x \#(x) = \#(r^{90c}(x)).$$

Notice in this truth the quantifier ranges over an infinite number of arrays; it will customarily be established via mathematical induction. But note as well that, courtesy of the imagistic cognition Penrose describes, this “infinitary” proposition can be instantly grasped. After all, just imagine the arrays growing incrementally in size, and imagine the rotation performed on these ever-increasing arrays. It’s obvious that rotation will preserve equality in each case, is it not? In Turing’s terms, intuition suffices to see what’s true, and ingenuity can be brought to bear to specify the (in this case tedious) inductive proof.

Here’s a second, more interesting example of mathematical thinking in *SOTM*, one involving the hexagonal numbers,

$$1, 7, 19, 37, 61, 91, 127, \dots$$

i.e., the numbers that can be arranged as ever-increasing hexagonal arrays (see Figure 1). Consider the cubes:

$$1 = 1^3, 8 = 2^3, 27 = 3^3, 64 = 4^3, 125 = 5^3, \dots$$

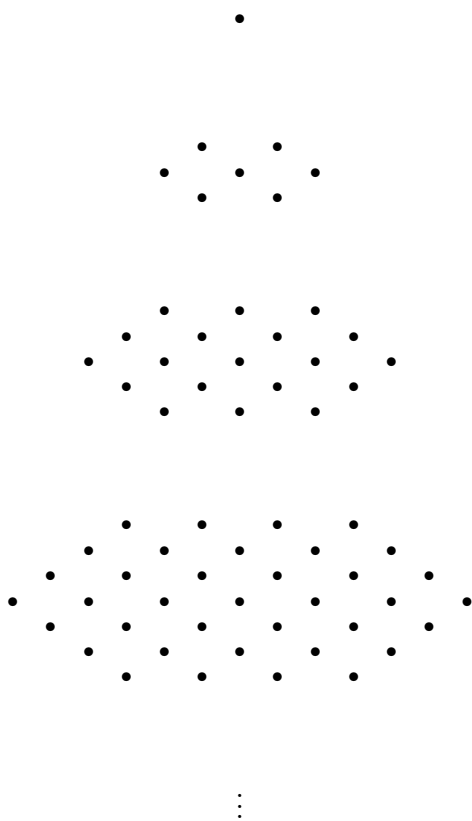
Let Turing machine $M_{\bar{e}}$ be defined as follows. $M_{\bar{e}}$ adds together the hexagonal numbers successively, starting with 1, checking to see if each sum is a cube. If so, the machine keeps working away; if not, it halts. Does $M_{\bar{e}}$ halt? No; that the pattern

$$1 = 1, 1 + 7 = 8, 1 + 7 + 19 = 27, 1 + 7 + 19 + 37 = 64, 1 + 7 + 19 + 37 + 61 = 125, \dots$$

continues forever is a conventionally established theorem. But as Penrose points out, this theorem can be grasped by a human mathematician courtesy of another, more elaborate imagistic trick. He writes:

⁵The array Penrose presents can be positioned on three tapes of a five-tape Turing machine, and the operation of rotation can result in three dots being placed contiguously on all five tapes. For a nice introduction to Turing machines with k tapes, see the textbook we make direct use of in section 4.1: (Lewis and Papadimitriou 1981). For a discussion of much more sophisticated image-processing AI systems than this hypothetical Turing machine, in the context of imagistic reasoning seemingly more exotic than what Penrose describes here, see (Bringsjord and Bringsjord 1996).

Figure 1: Hexagonal Numbers as Arrays



First of all, a cube is called a cube because it is a number that can be represented as a cubic array of points as depicted in Figure [2]. I want you to try to think of such an array as built up successively, starting at one corner and then adding a succession of three-faced arrangements each consisting of a back wall, side wall, and a ceiling, as depicted in Figure [3]. Now view this three-faced arrangement from a long way out, along the directions of the corner common to all three faces. What do we see? A *hexagon* as in Figure [4]. The marks that constitute these hexagons, successively increasing in size, when taken together, correspond to the marks that constitute the entire cube. This, then, establishes the fact that adding together successive hexagonal numbers, starting with 1, will always give a cube. Accordingly, we have indeed ascertained that $[M_c]$ will never stop. (Penrose 1994, p. 70)

We suspect that once again Penrose’s critics will almost by reflex maintain that the operations Penrose describes here are things that computers can be (and routinely are) programmed to do. This time the demonstration that the diagrams in question are S-D’s would be a tad more involved. We would need to represent, in FOL, three-dimensional grids and mechanical operations on them. Then a traditional-style proof would be crafted on the basis of these representations; once again mathematical induction would presumably be used. But such a proof is without question *not* what Penrose has in mind. He has in mind three-dimensional objects, objects that once again are operated upon and inspected, in thought-experimental fashion. The key is *seeing*, with one’s mind’s eye, that the three-faced objects, as they grow *ad infinitum*, will always be, from a certain perspective, hexagons. Once one visualizes the situation “stretching to infinity,” a conventional proof is superfluous, epistemically speaking. A journal may insist upon receiving the proof, but *you* know the truth without even bothering to write the first syllable of a conventional inference.

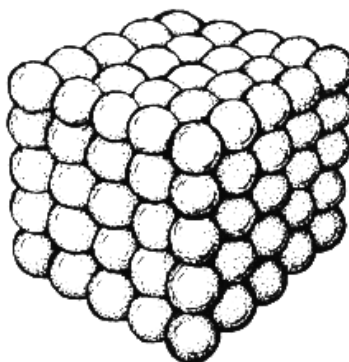


Figure 2: Cubic Array of Spheres

4 From Intuition to Ingenuity: How?

Of course, however sure we may be that Penrose has grasped the superiority of minds over machines when it comes to mathematical reasoning, the fact remains that you are unlikely to be convinced. There are two routes we can take to try to convince you. Though both of these routes are ones that occurred to us before meditating on Penrose’s examples and

accompanying intuitions, both, we now realize, can be viewed as running from his intuitions to third-person specification.

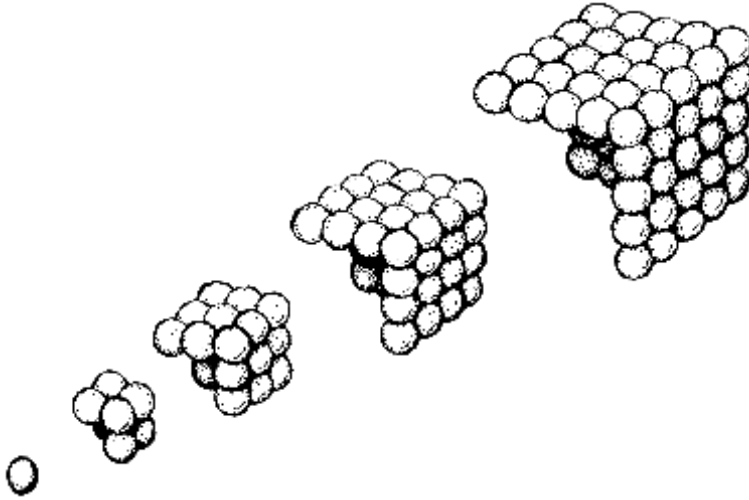


Figure 3: Each With a Back Wall, Side Wall, and Ceiling

In the first avenue, taken elsewhere by Bringsjord (1997), one first isolates and exploits mathematical reasoning that seems to be explicitly and irreducibly infinitary. The best example of such reasoning that we are aware of is found in *infinitary* mathematical logic.⁶

There is a second avenue available; it's the one we take herein. In broad strokes, the route runs like this. First, a formally valid argument surmounting problem 1 is formulated (our modal argument). Second, the key premise therein is defended by looking more closely at an aspect of mathematical cognition which, like the imagistic thinking Penrose has spot-

⁶The key idea is to find mathematical cognition that is provably *beyond* computation. Such cognition is exhibited by logicians and mathematicians who prove things in and about *infinitary* logics (which arose in no small part as a way to “surmount” Gödel I; see Barwise 1980 for a readable discussion of this point), such as the logic $\mathcal{L}_{\omega_1\omega}$. The basic idea behind $\mathcal{L}_{\omega_1\omega}$ is straightforward. This system allows for infinite disjunctions and conjunctions, and hence allows for infinitely long derivations, where these disjunctions, conjunctions, and proofs are no longer than the size of the set of natural numbers. (We use ω to denote the “size” of the set of natural numbers: the niceties of cardinal numbers needn’t detain us here.) Here is one simple formula in $\mathcal{L}_{\omega_1\omega}$ which is such that any interpretation that satisfies it is finite:

$$\bigvee_{n < \omega} \exists x_1 \dots \exists x_n \forall y (y = x_1 \vee \dots \vee y = x_n).$$

This formula is an infinite disjunction; each disjunct has a different value for n . One such disjunct is

$$\exists x_1 \exists x_2 \forall y (y = x_1 \vee y = x_2),$$

which says, put informally, there exist at most two things x_1 and x_2 with which everything in the domain is identical, or there are at most two things in the domain. It is a well-known fact that the proposition captured by this formula cannot be captured by a formula in a system at or below Turing machines. Since the behavior of some logicians and mathematicians centers around infinitary reasoning that can be accurately described only by formalisms that include such formulas (i.e., formalisms like $\mathcal{L}_{\omega_1\omega}$), computationalism is threatened. For the full argument see (Bringsjord 1997). A refined version of the argument can be found in (Bringsjord and Zenzen 2001).

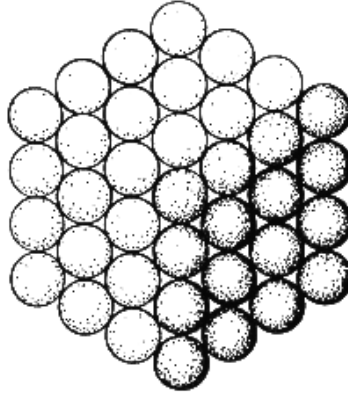


Figure 4: The Points Together Form a Hexagon

lighted for us, looks to be beyond the reach of computation. This aspect is the ability of a mathematician to understand through thought-experiments what's it's like to *very rapidly* experience some process or operation. We will explain this aspect of mathematical thinking momentarily, in connection with Turing's own thinking about the machines that now bear his name. But first, as planned, we turn to the modal refutation of computationalism.

4.1 The Modal Argument

Recall that Gödel's first incompleteness theorem can be expressed in "Turingish" terms as the fact that this problem is algorithmically unsolvable: Given an arbitrary Turing machine m and input u , does m halt on u ?⁷ Here is how this theorem, and a related one which in our experience a lot less people know about, are expressed on page 278 of the classic text *Elements of the Theory of Computation* by Lewis and Papadimitriou (1981):

Theorem 6.1.5

- (a) There is no algorithm that can determine, given a Turing machine m and an input string w , whether m accepts w .
- (b) For a certain fixed Turing machine m_0 , there is no algorithm that can determine, given an input string w , whether m_0 accepts w .

Intuitively, (a) says that there is no Turing machine that can "crack" every Turing machine (as to halting or non-halting); and (b) says that there is an impenetrable Turing machine: a machine that no other machine can crack. Notice that the sense of 'can' here, given that we are dealing with mathematical theorems, is very strong. Specifically, to say in this context that a Turing machine can't Φ is to say that it's logically impossible for the Turing machine to Φ .⁸ Accordingly, if we use the three-place predicate $Dmm'i$ for 'Turing

⁷Actually, as many forget, Turing (1936) was concerned with computable numbers; the halting problem, in the customary contemporary form we have just expressed it in, falls out as a consequence. We leave aside characterization of the relationship between Gödel I and Turing's (more general) results.

⁸Recall that algorithms and Turing machines are interchangeable by Church's Thesis. If you happen to be one of the few people on the planet who are skeptical about Church's Thesis, that's okay: simply start

machine m determines whether Turing machine m' halts on input i ,⁹ and if we assume to ease exposition a sorted calculus for Turing machines, inputs, and persons, we can symbolize (b) as⁹

$$(3) \forall m \exists i \neg \diamond D m m_0 i$$

Now suppose for *reductio* that computationalism is true, i.e., that persons are (physically realized) Turing machines, which can be symbolized as

$$(4) \forall p \exists m p = m$$

It follows from (3) and (4) that

$$(5) \forall p \exists i \neg \diamond D p m_0 i$$

However, given that there are persons, (5) is inconsistent with

$$(6) \forall p \forall i \diamond D p m_0 i$$

so indirect proof yields $\neg(3)$, that is, computationalism is false.

Two points should be made immediately about this argument (with others coming when we rebut objections).

First, it's important to grasp that the underlying modal argument isn't inseparably linked to a particular formal derivation. We have presented the derivation in a manner with which we happen to be particularly comfortable, but an infinite number of permutations are possible, once the core idea is grasped. The point here is important because the computationalist will not be able to dodge our modal argument by hiding behind such claims as that the Barcan Formula is invalid, or that identity statements cannot, in general, be necessitated. While there are formal versions of the argument that explicitly invoke BF and related aspects of quantified modal logic, we are in fact most strongly drawn to formal versions of the argument that stay within FOL, on the strength of innocent readings of "... can possibly determine whether ...". For example, the following two equations, where $D^\diamond xyx$ abbreviates the merging of \diamond with D to make an FOL-representable predicate meaning that x can (logically possibly) determine whether y halts on input z , are true, as can be verified by hand or by supplying them as input to a standard theorem prover:

$$\begin{aligned} & \{ \exists x (Mx \wedge \forall y (My \rightarrow \exists u (Iu \wedge \neg D^\diamond(y, x, u))), \forall x (Px \rightarrow \forall y \forall u ((My \wedge Iu) \rightarrow D^\diamond(x, y, u))), \\ & \quad \forall x (Px \rightarrow \exists y (My \wedge y = x)), \exists x Px \} \vdash \phi \wedge \neg \phi \end{aligned}$$

with a version of Theorem 6.1.5 in which 'algorithm' is supplanted with 'Turing machine.' For an argument against Church's Thesis, see "Chapter 5: A Narrative-Based Refutation of Church's Thesis" in (Bringsjord and Ferrucci 2000).

⁹Were we to make all the predicates explicit, with Mx for 'x is a Turing machine,' Px for 'x is a person,' and Ix for 'x is input for a Turing machine,' (3) would become either

$$(3') M m_0 \wedge \forall y (My \rightarrow \exists u (Iu \wedge \neg \diamond D(y, m_0, u)))$$

or

$$(3'') \exists x (Mx \wedge \forall y (My \rightarrow \exists u (Iu \wedge \neg \diamond D(y, x, u))))$$

$$\{Mm_0 \wedge \forall y(My \rightarrow \exists u(Iu \wedge \neg D^\diamond(y, m_0, u))), \forall x(Px \rightarrow \forall y\forall u((My \wedge Iu) \rightarrow D^\diamond(x, y, u))), \\ \forall x(Px \rightarrow \exists y(My \wedge y = x)), \exists xPx\} \vdash \phi \wedge \neg\phi$$

Here is the second point. Clearly, our underlying modal argument can be based on any number of related problems. Obviously, it can be based on Theorem 6.1.5(a). But it just as obviously can be based on many other problems. For example, we know that no Turing machine can determine whether an arbitrary arithmetical formula is true on the standard interpretation of arithmetic; and, again, this implies that no Turing machine can *possibly* make this determination. Yet it certainly seems logically possible for a person to make this determination. The same can be said for trying to determine, in generally, whether or not a first-order formula is a validity. And so on. However, we have for certain specific reasons based the above instantiation of the underlying modal argument on Theorem 6.1.5(b). These reasons pertain to the apparent fact that many logicians and mathematicians have a hard time accepting that *particular* problems are such that it's logically impossible for humanity to solve them. An impressionistic but seminal discussion of this matter can be found in (Minsky 1967). Minsky points out that while many are willing to accept that humanity can't possibly find a uniform procedure for solving an infinite class of problems, "Some mathematicians accept these uniform unsolvability results but prefer to believe that there can be no analogue for single problems" (Minsky 1967, p. 164). Gödel was an example of such a thinker. Though this may make a difference only to some readers, we point out that we herein chose a particular instantiation of the modal argument that appeals to a *particular* machine, m_0 . There are any number of other ways to "particularize" the modal argument.

5 Why the Key Premise is True

At first glance, this disproof is likely to look like a piece of legerdemain. After all, why is (6) true? Well, actually, the truth of (6) can be seen via two steps: by attending to the mathematics of information processing, and by unflinchingly attending to the underlying cognition that has *produced* this mathematics.

For the start of step one, note that there *are* information processing machines which can solve the halting problem — they just aren't *Turing* machines. There are in fact many machines that can exceed the "Turing Limit." Indeed, just as there are an infinite number of mathematical devices that are equivalent to Turing machines (first-order theorem provers, Register machines, the λ -calculus, abaci, . . . ; many of these are discussed in the context of an attempt to define computation in Bringsjord 1994), there are an infinite number of devices beyond the Turing Limit. As you might also guess, a small proper subset of these devices dominate the literature. In fact, three kinds of super-computational machines — analog chaotic neural nets, trial-and-error machines, and Zeus machines — are generally featured in the literature. In the interests of reaching a wider audience, we discuss only the last of these three devices here.¹⁰

¹⁰Analog chaotic neural nets are characterized by Siegelmann and Sontag (1994). For cognoscenti, analog chaotic neural nets are allowed to have irrational numbers for coefficients. For the uninitiated, analog chaotic neural nets are perhaps best explained by the "analog shift map," explicated by Siegelmann (1995),

Zeus machines are based on the character Zeus, described by Boolos and Jeffrey (1989). Zeus is a superhuman creature who can enumerate \mathbf{N} , the natural numbers, *in a finite amount of time*, in one second, in fact. He pulls this off by giving the first entry, 0, in $\frac{1}{2}$ second, the second entry, 1, in $\frac{1}{4}$ second, the third entry in $\frac{1}{8}$ second, the fourth in $\frac{1}{16}$ second, \dots , so that, indeed, when a second is done he has completely enumerated the natural numbers. Obviously, it's easy to adapt this scheme so as to produce a Zeus machine that can solve the halting problem: just imagine a machine which, when simulating an arbitrary Turing machine m operating on input u , does each step faster and faster \dots (There are countably many Turing machines, and those that don't halt are trapped in an unending sequence of the same cardinality as \mathbf{N} .) If, during this simulation, the Zeus machine finds that m halts on u , then a 1 is returned; otherwise 0 is given. Put in the abbreviatory first-order terms we availed ourselves of above, where z, z', \dots range over Zeus machines, we have

$$(7) \quad \forall z \forall i \diamond D z m_0 i$$

Now for step two, in which we will find that the mathematical cognition underlying the likes of (7) supports (6). How will this support be generated, exactly? From the logical point of view, we will see that the mathematical reasoning in question implies this conditional:

$$(8) \quad (7) \rightarrow (6)$$

Key proposition (6) will thus follow by *modus ponens* from (7) and (8).

To isolate the mathematical cognition we have in mind, return to certain aspects of Turing's original cognition when he was laying down the foundational concept of a Turing machine. Turing did *not* start with some mathematical description of a machine of some sort; he started with the concept of a *person* — or, to use his term, a ‘computist’ — carrying out primitive operations.¹¹ (Expressed in the key distinction of Turing's introduced above, Turing started with intuition and moved to ingenuity.) As Copeland, no doubt the thinker who knows the mind and history of Turing best, explains, what Turing imagined was “a human mathematician who is unaided by any machinery save paper and pencil, and who is

and summarized in (Bringsjord 1998). Trial-and-error machines have their roots in a paper by Hilary Putnam (1994), and one by Mark Gold (1994); both appeared in the same rather famous volume and issue of the *Journal of Symbolic Logic*. Trial-and-error machines have the architecture of Turing machines (read/write heads, tapes, a fixed and finite number of internal states), but produce output “in the limit” rather than giving one particular output and then halting. Here is a trial-and-error machine \mathcal{M} that solves the halting problem. Take some arbitrary Turing machine m with input u ; let $n^{m,u}$ be the Gödel number of the pair m, u ; place $n^{m,u}$ on \mathcal{M} 's tape. Now have \mathcal{M} print 0 immediately (for “No”), and then have it simulate the operation of m on u . If \mathcal{M} halts during the simulation, have it proceed to erase 0 in favor of 1 (for “Yes”), and then have it stop for good. It's as easy as that. For full exposition, along with arguments that human persons are trial-and-error machines, see (Kugel 1986), a seminal paper that situates trial-and-error machines nicely within both the formal context of the Arithmetic Hierarchy and the philosophical context of whether minds are computing machines.

¹¹In his inaugural writings on isomorphic points (independent, by the way, of Turing's), Post (1944) spoke of mindless “workers,” humans whose sole job was to slavishly follow explicit, excruciatingly simple instructions. Likewise, Charles Babbage modeled the calculating cogs in his Difference Engine (and, for that matter, in the never-built Analytical Engine) on an army of hairdressers hired by Baron Gaspard Riche de Prony, director of the École des Pont et Chaussées, to transform data tables into decimal form when France went decimal in 1799: see (Holt 2001).

working in accordance with ‘mechanical’ methods, which is to say, methods set out in the form of a finite number of exact instructions that call for no insight or ingenuity on the part of the person carrying them out” (1998*c*, p. 131). As Turing himself explicitly said: “A man provided with paper, pencil, and rubber, and subject to discipline, is in effect a universal [Turing] machine” (Turing 1969, p. 9).¹² It would seem that we can safely infer from such facts as that a Turing machine, given a natural number n , can determine whether n is even, that it is logically possible for a person to make the same determination. Likewise, it would seem that we can safely infer from the fact that a Zeus machine can determine whether a Turing machine m will halt on input i , that it’s logically possible for a person unaided by but (enough) paper and pencil to make the same determination — the person just has to work rather fast. In fact, we would be willing to wager that *you* apprehend the truth of (7), in the absence of all the mathematical details, precisely because you started to imagine the *person* Zeus described above doing what we said he does: he lists 1, then lists 2, then 3, and so on, moving faster at each step. Some of our readers, we are willing to bet, imagined *themselves* operating in this way.¹³ It should be clear that such mental maneuvers are similar in kind to the Penrosean ones discussed earlier in this paper.

It is important to realize that many other thinkers have independently affirmed proposition (6) on the strength of the technique of gedanken-experiment, in which they imagine what it’s like for a person to behave in the relevant manner. It is Bertrand Russell who seems to be have been the first to grasp the essence of Zeus machines. In a lecture in Boston in 1914 he said about Zeno’s paradox involving the race-course: “If half the course takes half a minute, and the next quarter takes a quarter of a minute, and so on, the whole course will take a minute” (Russell 1915, pp. 172–173). And later, when lampooning finitism as championed by Ambrose, Russell wrote:

Ambrose says it is *logically* impossible [for a man] to run through the whole expansion of π . I should have said it was *medically* impossible. . . . The opinion that the phrase ‘after an infinite number of operations’ is self-contradictory, seems scarcely correct. Might not a man’s skill increase so fast that he performed each operation in half the time required for its predecessor? In that case, the whole infinite series would take only twice as long as the first operation. (Russell 1936, pp. 143–144)

A number of other thinkers have conceived of that which our proposition (6) encapsulates. For example, Ralf Blake conceived this scenario quite a while ago,¹⁴ and more recently Salmon

¹²As Jack Copeland recently explained to us in conversation at a conference on Turing, Turing practiced what he preached here: Long before IBM’s Kasparov-beating *Deep Blue* came on the scene, there was Turing operating *qua* chess-playing machine! Before there was a computer to run programs on, Turing wrote a program for playing chess, and slavishly followed the program to play chess against a “normal” human opponent. The match was broadcast on the radio in England.

¹³It is interesting to note that Turing initially (1936) did not think it was *physically* possible for a Turing machine to be built (see Copeland 1998*c*, p. 130). It would be a mistake to infer from the mathematical fact that a Turing machine can determine whether an arbitrary natural number is even that it is physically possible for a person to make this determination on an arbitrary number.

¹⁴E.g., Ralf Blake wrote:

A process is perfectly conceivable, for example, such that at each stage of the process the addition of the next increment in the series $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$, etc., should take just half as long as the addition of the previous increment. But . . . then the addition of all the increments each to

(1975) and Copeland (1998a, 1998b) have imagined (and formalized) related “Zeusian” scenarios. Actually, the less colorful but more historically accurate name for the machines in question is “Weyl Machine,” for in 1927 Hermann Weyl considered a machine able to complete

an infinite sequence of distinct acts of decision within a finite time; say, by supplying the first result after $\frac{1}{2}$ minute, the second after another $\frac{1}{4}$ minute, the third $\frac{1}{8}$ minute later than the second, etc. In this way it would be possible . . . to achieve a traversal of all natural numbers and thereby a sure yes-or-no decision regarding any existential question about natural numbers. (Weyl 1949, p. 42)

Other names for Zeus machines are in the literature. For example, Copeland (1998a, 1998b) has recently referred to Zeus machines as ‘accelerated Turing machines’ and ‘super Turing machines.’

What about Gödel himself? And what about Turing? Well, Gödel seems to have been of the opinion that the human mind *in fact* enters an infinite number of distinct states. Turing explicitly argued for the view that the human mind is capable of only a finite number of distinct states (see pp. 92–93 of Wang 1974), and though Gödel rejected this argument because it presupposed a materialist conception of mind (Turing assumed that the mind equals brain; Gödel wrote and said on many occasions that the mind includes non-physical powers and parts), he also specifically wrote that “there is no reason why the number of states of the mind should not converge to infinity in the course of its development” (Wang 1974, pp. 325–326). It’s safe to say that Gödel, at least during certain stages of his life, would have regarded our (6) as obviously true.

Now, finally, to Turing. As we have just indicated, he believed that the mind was “capable” of only a finite number of states. But ‘capable,’ modally speaking, is slippery. Did Turing believe that it isn’t even *logically possible* for a person to perform as Zeus? Would he have rejected (6), if this proposition were presented to him independent of the argument in which we have situated it? Apparently not. In fact, it seems likely that Turing would have *affirmed* (6). The reason is bound up with the apparent fact that contemporary computer science embraces a mathematical scheme that well nigh entails (6), and with the fact that this scheme is due to none other than Turing himself. In his dissertation (Turing 1938–1939), Turing pondered the possibility of so-called *oracle machines*. These machines are architecturally identical to Turing machines, but are assumed to be augmented with an oracle which, upon being consulted about a Turing machine m and input i , returns a correct verdict as to whether m halts on i . Oracle machines are part of the mathematical canon of computer science today. For example, here is how a recently updated classic textbook on computability and uncomputability introduces oracles:

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:

each shows no sign whatever of taking forever. On the contrary, it becomes evident that it will all be accomplished within a certain definitely limited duration. . . . If, e.g., the first act . . . takes $\frac{1}{2}$ second, the next $\frac{1}{4}$ second, etc., the [process] will . . . be accomplished in precisely one second. (Blake 1926, pp. 650–651)

Suppose we were given a “black box” or, as one says, an *oracle*, which 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 . . . (Davis, Sigal & Weyuker 1994, p. 197)

More important than that Turing would probably have affirmed the likes of (6) is the fact that computer science itself commits to this proposition. As the quote from Davis et al. indicates, first oracles are imagined, not as mathematized, mechanical devices, but rather as intuitive objects. *After this*, oracles can be cashed out (as we’ve said) via Zeus machines, trial-and-error machines, analog chaotic neural nets, and so on. So here we see the same sequence: human imagination and intuition first, followed by formal ingenuity.

6 Objections

6.1 Objection 1: “You’re Misrepresenting Computationalism”

The first objection runs as follows: “You identify computationalism with the claim that people are Turing machines, but computationalists needn’t maintain any such thing. It suffices for the computationalist to claim that people can be *simulated* by computational systems (e.g., by Turing machines). After all, Turing’s test for machine mentality, now known simply as the ‘Turing Test,’ requires for a ‘passing grade’ that the computer in question be conversationally indistinguishable from a person; the test does *not* insist that a computer *be* a person, a bearer of such things as subjective awareness.”

This objection conflates “Strong” and “Weak” versions of computationalism and AI, and thus flies in the face of a fundamental distinction in AI and the philosophy of AI. The view that computing machines can simulate the behavior of persons is known as “Weak” AI; the view that people quite literally *are* computing machines, and hence that AI can eventually *replicate*, not just simulate, persons, is known as “Strong” AI. We aren’t attacking Weak AI. Doing so, for reasons one of us has explained elsewhere, would be foolish: Weak AI is almost certainly true.¹⁵ We seek to refute the kind of position nicely described by John 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*. (Haugeland 1981, p. 2)

While our (4) is a standard way to symbolize standard expressions of Strong computationalism, as in this one from Haugeland, we do of course realize that our way is one from among

¹⁵An explicit argument for the “obviousness” of Weak AI is given in (Bringsjord 2000).

many that might be chosen. An extensive list of candidate propositions for encapsulating the Strong computational conception of mind is presented in (Bringsjord and Zenzen 1997), with the “People are Turing machines” distillation chosen in the end, as accurate and fair. This distillation nicely captures the core of many statements of computationalism found in the literature, e.g., in (Simon 1980, Simon 1981, Dietrich 1990, Haugeland 1981, Newell 1980, Johnson-Laird 1988, Searle 1980, Barr 1983).

Nonetheless, perhaps Objection 1 can be sustained as follows: “Very well, I agree to talk not just of simulation, but replication; point well-taken. But I do think it makes a vast difference whether we say that persons are Turing machines or ‘real life’ embodied computers. Specifically, computationalism as the claim that people are physical computers is insulated from your modal argument; in fact, your argument is a non-starter once talk of Turing machines is supplanted with talk of such computers. To see this, note first that computationalism, encapsulated by your (4), now becomes, where c ranges over embodied computers,

$$(4') \quad \forall p \exists c p = c$$

Now, in order for there to be a parallel modal argument here, (3) must be replaced with

$$(3') \quad \forall c \exists i \neg \diamond Dcm_0i$$

But this proposition is false! Take a particular “real life” computer, say the Powerbook G4 on which you’re working at present; let’s dub it **G4**. Obviously, it’s logically possible for **G4** to crack m_0 : you’ve showed us how above. What **G4** needs to do is simply work in Zeusian fashion on m_0 and some input i . The scenario is easy to imagine — just as easy as it is to imagine that some *person* works in this fashion. Since (3') is false, the modal argument fails.”

The reply to this objection is straightforward. Computationalism is the view that people are *standard* computers — that is to say, computers that instantiate a formal scheme equivalent to that defined by Turing machines. Computationalism is *not* the view that people are hypercomputers, that is, information-processing devices capable of feats beyond the Turing Limit. After all, the holy grail for those thinkers seeking to overthrow computationalism has long been some cognition that exceeds the Turing Limit. Penrose, in this regard, is right on key: his is a search for mathematical cognition that involves information processing beyond the Turing Limit. In sum, then, the modal argument, as the previous objection makes plain, does not the slightest harm to the view that the mind is (at least in part) a hypercomputer. Indeed, this view is *defended* by one of us in a forthcoming book (Bringsjord and Zenzen 2001).

6.2 Objection 2: “But Then Computers Aren’t Computers!”

“Your argument can’t possibly be right. For if we know anything about computation, we know that Macintosh Powerbooks are physically instantiated Turing machines.¹⁶ But

¹⁶Actually, since Turing machines don’t have random access memory, this isn’t really true. It would be more accurate to say, e.g., that Powerbooks are physically instantiated register machines. A nice presentation of register machines can be found in (Ebbinghaus, Flum and Thomas 1984).

observe what happens when we replace reference to persons in your modal argument with reference to Powerbooks, a class we can assume b to range over. Proposition (4) becomes the fact that Powerbooks are Turing machines, that is,

$$(4'') \quad \forall b \exists m b = m$$

It now follows from (3) and (4'') that

$$(5') \quad \forall b \exists i \neg \diamond D b m_0 i$$

But this proposition, given that there are Powerbooks, is inconsistent with

$$(6') \quad \forall b \forall i \diamond D b m_0 i$$

It thus follows by indirect proof that Powerbooks aren't computers, which is patently absurd!"

This objection surreptitiously conflates two different senses in which Powerbooks (and indeed any class of physical computers) can be said to be Turing machines (or some other idealized type of machine). The first sense is an *architectural* sense only, while the second is an architectural sense *and* a temporal sense. For a physical machine to be a Turing machine (or some other idealized machine) merely architecturally means that the structures of the two correspond. Turing machines, for example, have read/write heads that move over tapes divided into squares, so it's possible to take a model railroad set and build a physical computer that "is" a Turing machine, by allowing the model railroad track to function as a tape, and so on. But this leaves the temporal issues completely open. The fact of the matter is that a Turing machine not only has a certain structure; it also has an inherent temporal limitation: viz., it can take only a finite number of steps in any finite time (as can be verified by looking at any relevant formalization; e.g., see again Lewis & Papadimitriou 1981.) The entire area of computational complexity in theoretical computer science is based upon this limitation. The problem is that (6') is true only if it construes Powerbooks as devices allowed to carry out an infinite number of operations in a finite time (in which case they aren't Turing machines, i.e., architecturally and temporally Turing machines), while (4'') is true only when Powerbooks are interpreted as *both* architecturally and temporally an instantiation of the Turing machine scheme. One traditional way to begin to mathematize those devices beyond the Turing limit, by starting with Turing machines, is to drop the inherent temporal limitations on Turing machines: this is exactly what Putnam (1994) and Gold (1994) did (see note 10).

6.3 Objection 3: "Zeus Merely *Seems* Coherent"

The third objection is inevitable: "You have said that the logical possibility of a Zeus machine solving the halting problem, which I agree to be a mathematical fact, commits one to the view that it's logically possible for a *person* to solve this problem by working faster and faster. But at most you have shown that it's apparently *coherent* that a person perform in this way."

For rebuttal we rely on a familiar point, one made recently by David Chalmers (1996), 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.¹⁷ Specifically, those who would resist us need to expose some contradiction or incoherence in the activity of Zeus, or the activity of persons as described by Russell, Weyl, Blake, Copeland, and so on. No such thing has ever been exposed.¹⁸

6.4 Objection 4: “As a Matter of Fact, Zeus is *Not* Coherent!”

The third objection goes like this: “As a matter of fact, some thinkers have argued that Zeus’ behavior is incoherent. These are people who don’t seem to share the attitude of Russell, Weyl, Blake, Copeland, and others you have conveniently listed. For example, A. W. Moore’s attitude about expanding π , expressed in connection with a form of Zeno’s paradox, runs rather counter to Russell’s:

Suppose that Achilles runs for half a minute, then pauses for half a minute, then runs for a quarter of a minute, then pauses for a quarter of a minute, and so on *ad infinitum*. At the end of two minutes he will have stopped and started in this way infinitely many times. Yet there is something repugnant about admitting this possibility, even as a conceptual — let alone a physical — possibility. For example, suppose that each time he pauses he performs a task of some kind, there being no limit to how quickly he can do this. Then at the end of two minutes he will have performed infinitely many of these tasks. He might, say, have written down the complete decimal expansion of π (3.141592...), for which he needs only a finite sheet of paper and the ability to write down digits that get smaller without limit (see [Figure 5]). We are loath to admit this as a conceptual possibility, though we seem bound to do so. (Moore 1990*a*, p. 4)

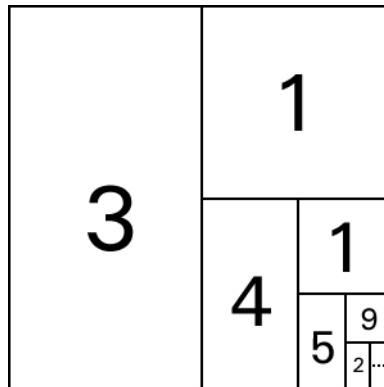


Figure 5: Expanding π as a Supertask

Moreover, such an expansion of π has been called in the literature a ‘super-task,’ and Moore (1990*a*) has provided a rather famous argument for the view that super-tasks are incoherent.”

Moore’s statement that “We are loath to admit this as a conceptual possibility” is mystifying. Who does the ‘we’ refer to? Not to the authors: Moore wrote *The Infinite* by himself.

¹⁷Chalmers 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.

¹⁸For more on these issues in connection with the computational conception of mind, see (Bringsjord 1999).

The idea must be that *thinkers in general* are reluctant to admit this conceptual possibility — but this just isn't the case. In our experience, most thinkers rendering a verdict on whether the expansion is a conceptual possibility, believe, with Russell and all the others, that it is. (It's also part of our experience that some of these thinkers lose this belief when apprised of our modal argument.)

We should probably mention that the notion of a limit, central to elementary calculus, by the lights of Salmon (1975) and others, presupposes the coherence of supertasks. Even children are frequently taught that supertasks are perfectly coherent, because they are prepared early on, in mathematics, for calculus down the road. For example, see Figure 6, which is taken from page 268 of (Eicholz, O'Daffer, Charles, Young, Barnett, Clemens, Gilmer, Reeves, Renfro, Thompson and Thornton 1995). Bringsjord's son, Alexander, in the 7th grade, was asked to determine the “percent pattern” of the outer square consumed by the ever-decreasing shaded squares. The pattern, obviously, starts at $\frac{1}{4}$, and then continues as $\frac{1}{16}, \frac{1}{64}, \frac{1}{256}, \dots$. When asked what percent “in the limit” the shaded square consumes of the original square, Alexander was expected to say “Zero” — but the notion of a limit was a bit tricky for him (perhaps understandably). When asked what percentage the shaded square would “get down to” if someone could work faster and faster, and smaller and smaller, at drawing the up-down and left-right lines that make each quartet of smaller squares, Alexander said zero. This is anecdotal, yes, but what it indicates is something in keeping with what we discussed earlier: some humans may start with an intuitive picture of supertask, and move from there to the formalisms in question (i.e., those seen in elementary calculus). It would be interesting to systematically poll students about these matters.

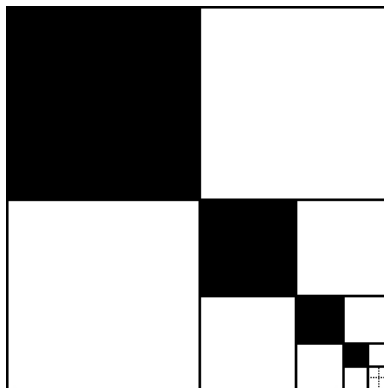


Figure 6: Picture of Supertask from Seventh Grade Math

At any rate, fortunately for Moore, as the end of Objection 4 intimates, his ammunition includes more than a suspicious straw poll, against which we could certainly pit our own. He does have one *argument* for the incoherence of super-tasks: it is part of a larger case for finitism (a view which denies even the existence of \mathbf{N} , a set we have of course explicitly invoked). Needless to say, we don't have the space to present and resolve one of the most profound issues in philosophy of mathematics, and if it takes an affirmation of finitism to ward off our modal argument, then our argument has all the force we could ever hope it to have. But we can focus specifically on Moore's argument for the incoherence of what the two of us, following Russell, Weyl, Blake, Copeland, and others, find transparently coherent.

Here, accordingly, is Moore's argument:

If it did make sense to say that I had just constructed all of the natural numbers in a minute, by the Zenonian procedure, then it would also make sense to say this: while I was constructing them, my constantly increasing speed of performance meant that time seemed to be going more and more slowly to me; it seemed that I was constructing them at a steady rate. Yet there is nothing that could count for me as a retrospective grasp of such an experience, in its apparent endlessness. (I could not have an apparently endless experience, apparently followed by further experience.) I must, subsequently, have forgotten all but an initial segment of it. How can this be? Surely what we have here is symptomatic of the fact that nothing could ever count, for anyone, as a grasp of an infinite reality. The grammar of 'infinity' is not geared to this. The special problems that arise when we envisage time seeming to go more slowly merely serve to make graphic an incoherence that is there to be acknowledged anyway — an incoherence that crept in at the very beginning of the story. It does not make sense to say that I have just performed infinitely many tasks of any kind, nor to say that anything is infinite in any respect. (Moore 1990*a*, p. 213)

Put in a more explicit form that can be evaluated, and ignoring the grandiose claim that finitism is vindicated, the argument runs like this:

- (9) If super-tasks are coherent, then it could seem to someone who (conceptually speaking) performs such a task that each action along the way took the same amount of time (say n seconds).
- (10) If it could seem to someone who performs a super-task that each action along the way took n seconds, then someone could retrospectively experience, or relive, each action for n seconds — and then proceed to have other, normal life experiences.
- (11) No one could retrospectively experience, or relive, each action in a super-task for n seconds — and then proceed to have other, normal life experiences.
- (12) Super-tasks are incoherent. [(9), (10), (11)]

Is this argument any good? Well, the argument appears to be formally valid, an instance of hypothetical syllogism and *modus tollens*; at any rate, we are prepared to concede that the argument is valid. Premise (11) certainly seems to be true. After all, $n + n + n + \dots$ will eventually exceed the amount of time a human has to operate with, whereas a super-task is based on some such sequence as $\frac{1}{2^n}, n = 1, 2, \dots$. Premise (9) seems extremely plausible. For suppose that Jones is a sprinter who runs 100 meter races in 9 seconds. Couldn't it nonetheless seem to Jones that he runs the race in 15 seconds? Couldn't it in fact seem to Jones that he runs the race in 30 seconds? And why not a minute? After all, we're talking here about Jones' subjective perspective. He could hallucinate during his sprint, or go into some kind of wild dream that seems to span 10 years. In general, then, when a human performs an action having property F , it may not seem to the human that that action has F ; the human may perceive the action to have a radically different property G .

So where does this leave the argument? It leaves it hinging on premise (10). But this premise is at best controversial; at worst, the premise begs the question against us. The problem is that it doesn't follow from the fact that a series of actions can *seem* to have

certain properties to someone, that that person can in any way genuinely experience actions having these properties. It may seem to me that a moment ago I jumped a tall building in a single bound (perhaps I had a dream), but it doesn't follow from this that I can experience jumping a tall building in a single bound. Likewise, it may seem to someone that they just squared the circle, but they cannot experience squaring the circle. (If one experiences squaring the circle, squaring the circle can be pulled off — but it *can't* be pulled off.)

Finally, why do we say that premise (10) may make the argument in question circular? We imagine that once one has (conceptually) performed a super-task, the only way to retrospectively experience this task is for the retrospective experience to itself be a super-task. To the extent that premise (10) rules out by fiat this method of retrospective experience, it preemptively rules on precisely what's at issue: Moore's argument becomes a *petitio*.

7 Conclusion

We don't claim to have provided a *disproof* of computationalism. We have given what seems to us to be a new and powerful modal argument against computationalism. Other objections are sure to be expressed, better versions of the modal argument are sure to be formulated,¹⁹ and only continued debate will determine whether we carry the day; that's philosophy. But we do claim to have recast the Gödelian case against computationalism, and to have thereby set the stage for a new phase in the debate.

¹⁹E.g., hypercomputational processes less exotic but more mathematically complicated than Zeussian ones will doubtless be deployed. Some of these processes can come from apparently naturally occurring physical processes, as in the bouncing of billiard balls among parabolic mirrors captured by a hypercomputational scheme known as the shift map (Siegelmann 1999, Moore 1990*b*, Bringsjord 1998). Since whatever is physically real (or even physically possible) is logically possible, and since bouncing billiard balls would seem to be harnessable by persons, it would seem logically possible for people to solve problems beyond the Turing Limit.

References

- Barr, A. (1983), Artificial intelligence: Cognition as computation, in F. Machlup, ed., ‘The Study of Information: Interdisciplinary Messages’, Wiley-Interscience, New York, NYK, pp. 237–262.
- Barwise, J. (1980), Infinitary logics, in E. Agazzi, ed., ‘Modern Logic: A Survey’, Reidel, Dordrecht, The Netherlands, pp. 93–112.
- Blake, R. (1926), ‘The paradox of temporal process’, *Journal of Philosophy* **23**, 645–654.
- Boolos, G. S. and Jeffrey, R. C. (1989), *Computability and Logic*, Cambridge University Press, Cambridge, UK.
- Bringsjord, S. (1992), *What Robots Can and Can’t Be*, Kluwer, Dordrecht, The Netherlands.
- Bringsjord, S. (1994), ‘Computation, among other things, is beneath us’, *Minds and Machines* **4.4**, 469–488.
- Bringsjord, S. (1997), An argument for the uncomputability of infinitary mathematical expertise, in P. Feltoich, K. Ford and P. Hayes, eds, ‘Expertise in Context’, AAAI Press, Menlo Park, CA, pp. 475–497.
- Bringsjord, S. (1998), Philosophy and ‘super’ computation, in J. Moor and T. Bynam, eds, ‘The Digital Phoenix: How Computers are Changing Philosophy’, Blackwell, Oxford, UK, pp. 231–252.
- Bringsjord, S. (1999), ‘The zombie attack on the computational conception of mind’, *Philosophy and Phenomenological Research* **59.1**, 41–69.
- Bringsjord, S. (2000), ‘Review of John Searle’s *The Mystery of Consciousness*’, *Minds and Machines* **10**(3), 457–459.
- Bringsjord, S. and Bringsjord, E. (1996), ‘The case against AI from imagistic expertise’, *Journal of Experimental and Theoretical Artificial Intelligence* **8**, 383–397.
- Bringsjord, S. and Ferrucci, D. (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Lawrence Erlbaum, Mahwah, NJ.
- Bringsjord, S. and Xiao, H. (2000), ‘A refutation of penrose’s gödelian case against artificial intelligence’, *Journal of Experimental and Theoretical Artificial Intelligence* **12**, 307–329.
- Bringsjord, S. and Zenzen, M. (1997), ‘Cognition is not computation: The argument from irreversibility?’, *Synthese* **113**, 285–320.
- Bringsjord, S. and Zenzen, M. (2001), *SuperMinds: A Defense of Uncomputable Cognition*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford, Oxford, UK.
- Copeland, B. J. (1998a), Even Turing machines can compute uncomputable functions, in J. Casti, ed., ‘Unconventional Models of Computation’, Springer-Verlag, London, UK, pp. 150–164.
- Copeland, B. J. (1998b), ‘Super Turing machines’, *Complexity* **4**, 30–32.

- Copeland, B. J. (1998*c*), ‘Turing’s O-machines, Searle, Penrose and the brain’, *Analysis* **58**(2), 128–138.
- Dietrich, E. (1990), ‘Computationalism’, *Social Epistemology* **4**(2), 135–154.
- Ebbinghaus, H. D., Flum, J. and Thomas, W. (1984), *Mathematical Logic*, Springer-Verlag, New York, NY.
- Eicholz, R. E., O’Daffer, P. G., Charles, R. I., Young, S. I., Barnett, C. S., Clemens, S. R., Gilmer, G. F., Reeves, A., Renfro, F. L., Thompson, M. M. and Thornton, C. A. (1995), *Grade 7 Addison-Wesley Mathematics*, Addison-Wesley, Reading, MA.
- Feferman, S. (1995), ‘Penrose’s godelian argument’, *Psyche* **2.1**. This is an electronic publication. It is available at <http://psyche.cs.monash.edu.au/psyche/volume2-1/psyche-95-2-7-shadows-5-feferman.html>.
- Gold, M. (1994), ‘Limiting recursion’, *Journal of Symbolic Logic* **30**(1), 28–47.
- Haugeland, J. (1981), *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge, MA.
- Holt, J. (2001), ‘The ada perplex: How byron’s daughter came to be celebrated as a cyber-visionary’, *The New Yorker* **March 5**, 88–93.
- Johnson-Laird, P. (1988), *The Computer and the Mind*, Harvard University Press, Cambridge, MA.
- Kugel, P. (1986), ‘Thinking may be more than computing’, *Cognition* **18**, 128–149.
- LaForte, G., Hayes, P. and Ford, K. (1998), ‘Why Gödel’s theorem cannot refute computationalism’, *Artificial Intelligence* **104**, 265–286.
- Lewis, H. and Papadimitriou, C. (1981), *Elements of the Theory of Computation*, Prentice Hall, Englewood Cliffs, NJ.
- Lucas, J. R. (1964), Minds, machines, and Gödel, in A. R. Anderson, ed., ‘Minds and Machines’, Prentice-Hall, Englewood Cliffs, NJ, pp. 43–59. Lucas’ paper is available online at <http://users.ox.ac.uk/~jrlucas/mmg.html>.
- Minsky, M. (1967), *Computation: Finite and Infinite Machines*, Prentice-Hall, Englewood Cliffs, NJ.
- Moore, A. W. (1990*a*), *The Infinite*, Routledge, New York, NY.
- Moore, C. (1990*b*), ‘Unpredictability and undecidability in dynamical systems’, *Physical Review Letters* **64**, 2354–2357.
- Newell, A. (1980), ‘Physical symbol systems’, *Cognitive Science* **4**, 135–183.
- Penrose, R. (1989), *The Emperor’s New Mind*, Oxford, Oxford, UK.
- Penrose, R. (1994), *Shadows of the Mind*, Oxford, Oxford, UK.
- Penrose, R. (1996), ‘Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind*’, *Psyche* **2.3**. This is an electronic publication. It is available at <http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html>.

- Post, E. (1944), ‘Recursively enumerable sets of positive integers and their decision problems’, *Bulletin of the American Mathematical Society* **50**, 284–316.
- Putnam, H. (1994), ‘Trial and error predicates and a solution to a problem of mostowski’, *Journal of Symbolic Logic* **30**(1), 49–57.
- Russell, B. (1915), *Our Knowledge of the External World as a Field for Scientific Method in Philosophy*, Open Court, Chicago, IL.
- Russell, B. (1936), ‘The limits of empiricism’, *Proceedings of the Aristotelian Society* **36**, 131–150.
- Russell, S. and Norvig, P. (1994), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Saddle River, NJ.
- Salmon, W. C. (1975), *Space, Time and Motion: A Philosophical Introduction*, Dickenson, Encino, CA.
- Searle, J. (1980), ‘Minds, brains and programs’, *Behavioral and Brain Sciences* **3**, 417–424. This paper is available online at <http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html>.
- Siegelmann, H. (1995), ‘Computation beyond the turing limit’, *Science* **268**, 545–548.
- Siegelmann, H. and Sontag, E. (1994), ‘Analog computation via neural nets’, *Theoretical Computer Science* **131**, 331–360.
- Siegelmann, H. T. (1999), *Neural Networks and Analog Computation: Beyond the Turing Limit*, Birkhäuser, Boston, MA.
- Simon, H. (1980), ‘Cognitive science: The newest science of the artificial’, *Cognitive Science* **4**, 33–56.
- Simon, H. (1981), ‘Study of human intelligence by creating artificial intelligence’, *American Scientist* **69**(3), 300–309.
- Turing, A. (1938), *Dissertation for the PhD: “Systems of Logic Based on Ordinals”*, Princeton University, Princeton, NJ.
- Turing, A. (1939), ‘Systems of logic based on ordinals’, *Proceedings of the London Mathematical Society (series 2)* **45**, 161–228.
- Turing, A. (1969), Intelligent machinery, in B. Meltzer and D. Michie, eds, ‘Machine Intelligence’, Edinburgh University Press, Edinburgh, UK, pp. 1–24.
- Turing, A. M. (1936), ‘On computable numbers with applications to the entscheidung-problem’, *Proceedings of the London Mathematical Society* **42**, 230–265.
- Wang, H. (1974), *From Mathematics to Philosophy*, Keagan Paul, London, UK.
- Weyl, H. (1949), *Philosophy of Mathematics and Natural Science*, Princeton University Press, Princeton, NJ.