The Case Against Computational Theory of the Mind:
A Refutation of Mathematically-Contingent Weak A.I.

Anthony D. Rhodes
Portland State University, 2011

"The Whole Dignity of Man Lies in His Thought"
– Pascal

In the last century, the debate over the nature of the mind has enjoyed a renewed prominence and relevance. Gregory Chaitin has, for instance, recently made the bold claim that the pure *Gedankenexperiments* provided by philosophers of mind in the 20th century directly engendered one of mankind's crowning intellectual achievements: the invention of the computer.[1] For the purposes of this paper, my interests lie in assessing and critiquing the current debate centering around the question prompted by the aforementioned genealogy of ideas, namely: *is the human mind a computer*? It is likely that the answer to this question will have an important impact upon considerations in the philosophy of mind as well as conceptions of the nature of the self over the course of the next century.

I will use an article published by the preeminent philosopher of mind, John Searle, entitled "Roger Penrose, Kurt Gödel, and Cytoskeletons" (1997) as my main point-of-entry into the "computational theory of mind" (i.e. the theory that human minds are essentially *equivalent* to *computational machines*) discussion. Searle is well-known for his attempted refutation of the core premises of the computational theory of the mind, known as the "Chinese Room" thought experiment (c. 1980). In the present paper I wish to focus on arguments that spawn from the theses of the Chinese Room argument, particularly with regard to Searle's more recent musings on Penrose and Gödel. Using concrete mathematical and logico-deductive tools, Searle's work on Penrose and Gödel frames the computational theory of the mind problem more cogently than many of the more theoretically informal or purely speculatively studies on this topic. The tropism toward a disciplinary inclusiveness with regard to investigations into questions in the philosophy of

mind is an inevitable and intellectually profitable trend. Many of the most favorable findings in recent decades in this field have come out of studies which chart the intersection of cognitive science, computer science, mathematics, neuroscience, psychology and philosophy – and Penrose is undoubtedly one of the foremost interdisciplinarians active in the philosophy of mind.

The origins of the *man vs. machine* debate for the philosophy of mind have their genesis, in part, in the "millennium problems" put forward by the eminent mathematician David Hilbert in 1900. Hilbert, who commanded a wide net of influence within the philosophy of mathematics in the first part of the 20th century, was committed to a type of philosophical *formalism.* Building upon the pioneering work of logicians like Frege and Dedekind, Hilbert sought to ground mathematics with a formal system of logic which derives results beginning from basic axioms using finitistic inferential rules (*à la* a Euclidean methodology). The goal of formalism is to demonstrate the internal consistency – and hence, the theoretical soundness – of mathematics. Despite the seeming persuasiveness of this program, the vexing findings of Russell, Cantor and above all Gödel in the first half of the century eventually proved that the agenda of formalism was, if not entirely nugatory, at the very least hopelessly unrealizable. These insuperable setbacks did not, however, put logical formalism to rest. On the contrary, as much as Gödelian Incompleteness rendered the grandiose epistemic aspirations of Science writ large, *viz*. mathematics, unworkable, formalism or variants of it have nevertheless exerted a persistent and polyvalent influence on discourses in the philosophy of science.

In one sense, we moderns are the direct beneficiaries of such historical "dismissals" of the seeming severity of Incompleteness. The specter of Incompleteness has, curiously,

led in a direct way to the reinvigoration of the study of formal systems of logic in the latter half of the 20th century. These lines of inquiry have proven fruitful in expanding conceptions of scientific "systems," as well as paving the way for additional deep-seated enigmas. In what follows, I wish to reconsider the philosophical implications of Gödel's findings with respect to the question of whether the human mind is in fact a computer. I will demonstrate how a version of Incompleteness may be used to bolster the notion of the inimitability of the human mind by refuting the "weakest" version of computational theory of the mind (CTM).[2] I begin this approach with a discussion of Turing.

Although he is unanimously considered one of the greatest scientific minds of the past century, Alan Turing is rarely given appropriate recognition for his important contributions to philosophy.[3] Contemporary literature has reassessed the importance of Turing's philosophical investigations, particularly in relation to the mind/computer problem.[4] Frequently, Turing is presumed (unjustly, in my view) to be a proponent of the CTM. This belief is largely informed by the sanguine predictions which Turing puts forward in his famous "Computing Machinery and Intelligence" (1950) paper where he conjectures that computational simulations of human minds could exist by the close of the 20th century (a prediction that he based on an assurance that storage capacities and processing speed limitations would not, alone, obviate the development of such a "perfect" simulation). But as we shall see, Turing's own *constructivist* reworking of Gödel's Incompleteness leads, conversely, through subsequent results by the logician John R. Lucas and Penrose, to a repudiation of CTM. Turing's philosophical inquiries by way of the inventions of both: (1) Turing machines and (2) the Turing test/the "Halting problem" are crucial to these explorations and require further examination.

Turing briefly described what has since become known as a "Turing machine" in a thought experiment from a paper dating back to 1937, and his more refined definition – a definition which essentially inaugurated the modern field of artificial intelligence (A.I.) – first surfaced in a 1948 essay entitled "Intelligent Machinery." Turing describes a machine consisting of several components: a read/write head, an infinite capacity/infinite length tape marked out into squares, and a *finite* table that suffices to define the internal instructions (read: axioms) of the machine/program. Typically, one can describe a Turing machine in terms of ordered *4-tuples*, e.g. (q_1, S_r, O_s, q_2). These *4-tuples* can be interpreted as follows: When the machine is in state q_1 and its read/write head "reads" symbol S_r (for the sake of simplicity, the machine-lexicon may consist solely of binary symbols *0* and *1*) from the table of instructions, then the machine proceeds to implement operation O_s, where the permissible operations include basic processes such as: move head left/right, write/erase, and so forth; after the execution of the designated operation, the machine transitions to the specified "new" state, q_2. It is not difficult to see, for example, that a machine constructed in this way can successfully add numbers together (and by extension, perform any arithmetic operation you like). I illustrate this notion with a common textbook example of a Turing machine that adds any two arbitrary numbers together.

$$\begin{cases} q_0\ 1\ B\ q_0 \\ q_0\ B\ R\ q_1 \\ q_1\ 1\ R\ q_1 \\ q_1\ 0\ 1\ q_2 \\ q_2\ q\ R\ q_2 \\ q_2\ B\ L\ q_3 \\ q_3\ 1\ B\ q_3 \end{cases}$$ "Turing addition"

This machine adds numbers *m* and *n* in the following fashion: On its tape the numbers *m* (a

string of 1's of length *m*) and *n* (a string of 1's of length *n*) separated by a zero are inscribed. If we assume the machine is initialized in the state q_o, its read-head scanning the leftmost 1 of the *m*-string, then it will add the two strings by deleting the separating zero and moving the entire *n*-string to the left one position. The end result is a continuous string of 1's the length of which is $m + n$.[5]

Such a machine is an effective *computational machine*, which is to say that any theoretical computation (even the more odious, infinite type) can be envisaged by such a machine. The remaining qualification which *universalizes* such a machine was implemented by Alonzo Church in the 1940s. Church used lambda-calculus to produce a more mathematically rigorous sense of a *Universal Turing Machine* (UTM), a Turing machine capable of simulating any other Turing machine by means of various embeddings/encoding schemas. The Church-Turing thesis states that "everything computable is computable by a Turing machine" and is considered the theoretical consummation of these developments.

It is incumbent that we understand how the Turing machine thought experiment informs Searle's rebuttal of attempts to equate mind and computer. Although my analysis is congruent with Searle insofar as we both agree that computers aren't minds (and vice versa), as he argues with his "Chinese Room" experiment, I nevertheless disagree with certain aspects of Searle's case against CTM. I will, to this end, attempt to bolster the line of argumentation adopted by Penrose in building a case against weak A.I.

In his book *Shadows of the Mind* (1994), Penrose identifies four possible forms that a mind-computer isomorphism could conceivably take, the first two of which are commonplace designations. In the first case, *strong A.I.* encompasses the position that

consciousness and other mental phenomena consist entirely in computational processes. *Weak A.I.*, conversely, requires that brain processes cause consciousness, and these processes can be potentially simulated on a computer. The remaining positions, the third of which Penrose endorses, consist in the notion that brain processes cause consciousness but these processes "cannot even be properly simulated computationally", while the last position (which I do no pursue in this paper) alleges that science cannot explain consciousness.

Before I unpack Searle's position (he concedes the legitimacy of weak A.I. but not strong A.I.) I want to first clarify the manner in which I propose to appraise CTM from an ontological point of view. After all, if CTM suggests that the mind and computer are *equivalent*, what here is meant by equivalent? Leibniz' principle of *the identity of indiscernibles* provides a sound way to check the purported equivalence of mind and computer upheld by CTM. This criterion is arguably the most widely-accepted principle relating to the assignation of ontological equivalence. The identity of indiscernibles states that two entities are identical if they have all *conceivable* (which is to say not merely empirical) properties in common. Stated more formally, then: if, for every property F, object x has F if and only if object y has F, then x is identical to y; alternatively, in the language of symbolic logic we have: $\forall F(Fx \leftrightarrow Fy) \rightarrow x = y$. Let us now apply this analytic standard to the claims made by CTM in order to illuminate Searle's approach, beginning with the famous Chinese Room argument.

Searle presents his arguments in the article "Minds, Brains and Programs" (1980). From the outset, he makes clear his position against weak A.I.: "my discussion [is] directed [against the specific] claim that the appropriately programmed computer literally has

cognitive states and that the programs thereby explain human cognition."[6] Searle draws an analogy between computer simulations of human mental phenomena and the process of mindless symbol shunting. The thought experiment is described as follows: Suppose that a monoglot, call him Searle, who speaks only English and understands not a word of Chinese is placed in a room. In the room Searle has access to a large instructional table of conditional statements which direct Searle to reply to such and such Chinese symbol with such and such intelligible response (assume that the table is exhaustive so that Searle is prepared to handle any string of questions written in Chinese). Questioners from outside the room pass Searle a series of yes/no queries written in Chinese. Although such questions are inscrutable to Searle, he nonetheless uses the table to formulate coherent replies to each of the questions posed to him so that his interlocutors are unable to distinguish his inability to read and write in Chinese. They therefore conclude that Searle *understands* Chinese; Searle *qua* a Turing machine (note that the table serves as a *program* or a set of axioms if one extends the analogy) has passed the Turing test; *mutatis mutandis*, the computer and mind are therefore equivalent.

But does such a situation pass muster with respect to our standard-bearing test for ontological indiserniblity? Searle reckons not. Recall that by Leibniz' lights, the mind and computer are equivalent if and only if every conceivable property that holds for the mind is preserved for the computer. Searle alleges that the inhabitant of the Chinese Room does not *understand* Chinese in the same way that a native speaker does; moreover, the process of mindless symbol shunting in no way helps to explicate the internal cognitive states of the native speaker. Ergo:

> Whatever purely formal principles you put into [a] computer, they will not be sufficient for understanding, since a human will be able to follow the

> formal principles without understanding anything. No reason whatever has been offered to suppose that such principles are necessary or even contributory, since no reason has been give to suppose that when I understand English I am operating with any formal program at all.[7]

Whereas consciousness is intrinsic to the human mind, computers *qua* Turing machines are bereft of such a quality. It stands to reason that computers are *not* minds, or so Searle's argument goes. Intentionality issues aside, the distinction between *syntactic understanding* and *symbolic understanding* lies at the heart of Searle's discussion. The computer, *viz*., Searle in the Chinese Room, analyzes its various data inputs solely on the basis of symbolic scrutiny: $Symbol:\Delta \overset{invokes}{\rightarrow} Operation:\Phi(\Delta)$, $Symbol:\sqcup \overset{invokes}{\rightarrow} Operation:\Phi(\sqcup)$ and so on. Human beings conversely understand these symbols within a broader semantic context, that is, from an "external point of view." Searle remarks conclusively: "syntax is insufficient for semantics."

While many may find the core of Searle's hypothesis against A.I. convincing, as I do, the Chinese Room example has nonetheless proved to be notoriously susceptible to a wide variety of counterarguments. To his credit, Searle anticipated a great many of these critical replies. In the end, Searle's general retort involves a basic claim about intentionality. No matter the elaborateness of the computer script or whether the computer in some fashion internalizes its program in toto (see the *systems reply*), or whether we allow for the possibility that a sentient being inhabit and thereby control a physical embodiment of a computer program (see *the robot reply*), the formal system itself never possesses an innate *intentionality.* Certainly the programmer of such a computer exudes a demonstrable intentionality. However, this intentionality is not carried over into the program with respect to the program's causal powers nor with respect to the "content" of the computational states induced by the program, as these states are entirely formal and

hence content-less.

In building my own response to Searle, I wish to avoid many of the eristic concerns that attend the rebuttal of strong A.I. in the form of the Chinese Room argument. The key to such a tactic involves an investigation into the unsolvability of the Halting problem which Turing presented in 1936 as a constructive or computational analogue to Gödel's 1931 Incompleteness Theorem.

The rudiments of the unsolvability of the Halting problem as it applies to a rejection of CTM are as follows. Naively, one can imagine certain elementary problems in mathematics whose truth or falsity can be known to us *semantically* or perhaps better put: structurally, and yet a purely computational approach to this same problem yields an inconclusive result. A trivial example might involve, for instance, the case in which we ask whether the sum of some pair of even numbers ever produces an odd number. Any schoolchild knows the falsity of such a claim because they can *see* the result, and this is precisely the point. More formally, we can concisely prove the result as follows:

$$2m + 2n = 2(m + n) \quad \forall m, n \in \mathbf{Z}$$

So the sum of any two even integers is even; *a fortiori*, an even number cannot be odd. Now if a computer exhaustively attempted to discover whether there are two even numbers whose sum is odd, e.g., "2+2 is even [next]…2+4 is even [next]…2+6 is even [next]…" such a procedure would never halt and therefore the answer to this problem would remain, in a computational sense, unknown. But the astute reader will, in all likelihood, find quibbles with this particular example. Isn't the statement: the sum of two even numbers is _____, simply, *pace* Kant, an *analytic a priori* proposition – whereupon the evenness of the sum of two numbers is connoted simply through the meaning of the term even itself?

Certainly such a criticism is perfectly valid. Yet things aren't quite as simple if we devise a more intricate problem.

Take as a second example a question regarding prime numbers. Bear in mind that a prime numbers is any integer, call it *p,* that is greater than or equal to two and whose only positive divisors are *p,* itself, and the number one. Now prime numbers are important for all sorts of reasons – but at this juncture let us simply ask, as many have wondered before: *are there an infinite number of primes*? The answer, which, to the best of anyone's knowledge first dates to Euclid, is a resounding yes. Notice that this result is less apparent than the previous example; we might, in the language of Kant, term it a *synthetic a priori* proposition. The idea is that it takes a bit more work in order to *see* the truth in the assertion that there are an infinite number of primes. Euclid proposed the following argument:

> *Proof:* Suppose not, and allow that there are only a finite number of primes: $\{p_1,p_2,p_3,\ldots,p_n\}$. Define the number $p^* = p_1p_2p_3\ldots p_n + 1$, which is to say p^* equals the product of all primes plus one. Then we are faced with two possibilities. Either p^* is prime, in which case our original finite set of primes is incomplete, a contradiction; otherwise, p^* is composite (non-prime) whereby the *fundamental theorem of algebra* guarantees that it is divisible by some prime number – but this too is a contradiction because p^* is not divisible by any prime by its very construction. Consequently, by the method of *reductio ad absurdum,* there exist an infinite number of primes.□

Just as in the previous example with even numbers, the truth or falsity of a mathematical proposition may require extrinsic or special structural knowledge of the given elements at hand in addition to information about the relational properties of these entities in order to ascertain a definitive result. With the prime number example, we are, I think, advancing closer to the line of demarcation between what is mentally possible and that which is computationally impossible; since this problem, for instance, offers no simple

computational solution. This disjunction underpins the basic strategy employed by Penrose in rejecting CTM through a refutation of weak A.I. Thus, by applying Leibniz' concept of identity from before, if one can demonstrate that *x* (the mind) possesses property *F* and *y* (a computer/Turing machine) lacks *F*, then *x* and *y* are dissimilar; hence minds and computers are not equivalent. Interestingly, if one could similarly show that the *converse* holds, that is, if there are properties intrinsic to a computer (read: things "known" by a computer) that a mind does not possess (read: things that we cannot know) then CTM fails, analogously (I explore these possibilities below).

Following Penrose's arguments, we are now ready to properly generalize the mind/computer disjunction sketched above by investigating, the results of Gödel as well as Turing's correlated formulation of the Halting problem. Gödel's arguments for Incompleteness incorporate a form of the well-known Liar's Paradox ("this statement is false") to make claims about provability within finistic formal systems, e.g. Peano arithmetic (PA). Through the use of an ingenious coding scheme, Gödel devises a mapping that assigns propositions in the given formal system to various natural numbers. Equipped with the arithmetical, or some such, structure that inheres in this mapping, Gödel subsequently reformulates the Liar Paradox as a *metamathematical* or self-referential statement about provability, *viz*., *this statement is unprovable in PA*. The further catch, which requires a good amount of formal machinery to fully illustrate, is that such a "Gödel statement" is in fact known to be *true* outside of the compass of the formal system at hand. This result is often characterized as the main outcome of Incompleteness.

In contrast to the non-constructive nature of Gödel's proof (it is, for instance, hard to wrap one's hands around a Gödel sentence in practice), Turing's reworking in the form

of the Halting problem yields a somewhat more practical result. The Halting problem poses the question as to whether there is a general algorithm that can accurately determine, given a description of a specific program, whether this program will eventually halt or continue to run forever. Turing proved that there is no such algorithm that can account for any arbitrary program and its associated input values. Obviously, many programs are evidently terminable; and these programs are therefore of little concern. However, let us reconsider the prior example regarding prime numbers. Now perhaps there is a simple algorithm we could construct to test this specific case; the algorithm might go something like this: "IF program requests search for maximal prime THEN return (0)" (where 0 denotes that the program is interminable). But now consider the program that asks a computer to find an odd perfect number (a currently unsolved problem) or to find a counterexample to the Goldbach conjecture (also unsolved). We can get a sense perhaps, that constructing a universal algorithm to test for terminability is deeply problematic.

Penrose's idea is that by extension of Incompleteness through the unsolvability of the Halting problem, it is possible to construct a computation that we can *see* does not ever stop and yet no universal algorithm can tell us this. But this result holds for any set of "computational algorithms"; whereupon he concludes that we are not computers carrying out an algorithm.

Before assessing this conclusion, I would like to walk through Searle's own version of Penrose's take on Turing and Gödel for two chief reasons. In the first case, his exposition is the most concise recounting of the proof of the unsolvability of the Halting problem that I have encountered; and secondly, the proof itself uses a version of a technique commonly known as *Cantor's diagolization argument*, which I will later

replicate during the brief discussion of consciousness contained at the end of this paper.

Proof (sketch): For any number n we can think of computational procedures C_1, C_2, C_3, etc., on n as dividing into two kinds, those that stop and those that do not stop. Now how can we find out which procedures never stop? Well suppose we had another procedure (or finite set of procedures) A which when it stops would tell us that the procedure $C(n)$ does not stop. Think of A as the sum total of all the *knowable and sound* (read: algorithmically correct) methods for deciding when computational procedures stop. So if the procedure A stops then $C(n)$ does not stop. Now think of a sequence of well-ordered computations numbered $C_1(n)$, $C_2(n)$, $C_3(n)$, etc. These are all the possible computations that can be performed on n. These would include basic arithmetic and algebraic properties such as multiplying a number by n, squaring n., etc. Since we have numbered all the possible computations on n we can think of A as a computational procedure which given any two numbers q and n tries to determine whether C$q(n)$ never stops. Suppose, for example, that $q = 17$ and $n = 8$. Then the task for A is to figure out whether the 17^{th} computation on 8 stops. Thus, if $A(q,n)$ stops, then $Cq(n)$ does not stop.

Previously we stipulated that the sequence $C_1(n)$, $C_2(n)$,… included *all* the computations on n, so for any n, $A(n,n)$ has to be a member of the sequence $\{C_n(n)\}$. Well, suppose that $A(n,n)$ is the kth computation on n, that is, suppose

$$A(n,n) = C_k(n). \quad (*)$$

Now consider the case when $n = k$, so that $A(k,k) = C_k(k)$. From above, it follows that if $A(k,k)$ stops, then $C_k(k)$ does not stop. However, if we substitute the identity into (*), we have: if $C_k(k)$ stops, then $C_k(k)$ does not stop. But if a proposition implies its own negation, it is false. Thus: $C_k(k)$ does not stop.

It therefore follows that $A(k,k)$ does not stop either, because it is the same computation as $C_k(k)$. This indicates that our comprehensive set of procedures is insufficient to tell us that $C_k(k)$ does not stop, despite the fact that we know it stops.

So A can't tell us what we really know, namely that $C_k(k)$ does not stop. Thus from the knowledge that A is sound, we can show that there are some nonstopping computational procedures, such as $C_k(k)$, that cannot be shown to be nonstopping by A. So we know something that A cannot tell us, so A is not sufficient to express our understanding. But A included all the knowably sound algorithms we had. Thus no knowably sound set of computational procedures such as A can ever be sufficient to determine that computations do not stop, because there are some such as $C_k(k)$ that they cannot capture. So we are not using a knowably sound algorithm to ascertain what it is that we know. [8]

Ergo, we are not computers.□

Whereas I consider the above argument basically a sufficient refutation of weak A.I., and accordingly a refutation of CTM, Searle conversely finds this reasoning deficient for the former purpose in one significant respect. Searle admits that Penrose has shown that the mind cannot be simulated at the level of mathematical reasoning by a program that uses only sound methods of mathematical proof. However, Searle objects to Penrose's broader conclusion that the mind "*cannot be simulated under any description whatever*." I view this characterization as something of a straw man. Penrose seems to say, rather, that it is impossible to adequately simulate our minds, and in particular our mathematical or logico-deductive mental capacities, by any *mathematically-contingent* A.I.

As I understand it, there are two main points to Searle's counterargument. In the first case, Searle attempts to exploit the *multiple realizability* issues that linger beneath the surface of Penrose's thesis. Searle asserts:

> [From Penrose's arguments] it does not follow that there cannot be computational simulation of the very same sequence of events at the level of brain processes, and we know that there must be such a level of brain processes because we know that any mathematical reasoning must be realized in the brain. Not only all neurobiologists but Penrose himself would agree that any change in mental states must be perfectly matched by a change in brain states.[9]

In truth, Penrose does not outright deny the possibility of the multiple realizability of particular mental states. Certainly, in theory, the mental state: "performing addition", might be realized equally by a bundle of neurons or some electrical impulses sent through a tube sock (electrical conductivity issues notwithstanding). Searle's fallacy however runs deeper in this instance. The claim that "all neurobiologists and Penrose would agree that change in mental state must be

perfectly matched by a change in brain states" commits Searle to an irrevocable contradiction. Remember that Searle agrees, along with Penrose, that strong A.I. is infeasible, so that he claims that mental phenomena in total cannot be properly simulated by a computer. But Searle seems to suggest in the preceding comment that there can be a perfect, isomorphic mapping between mental states and brain states. This belief is incompatible with a rejection of strong A.I. If the hypothesis of strong A.I. is false, then brain states are not equivalent to mental states (for if they were, an ideal physical model of the brain would satisfy the conditions of strong A.I.); and if brain states are different from mental states then Searle is a committed *dualist* and his statement above is nonsensical.

I likewise reject Searle's second rejoinder to Penrose. Searle explains his reasoning:

> He [Penrose] thinks he has shown that you could not program a robot to do all the mathematical reasoning that human beings are capable of. But once again that claim has to be qualified. He has, if he is right, shown that you could not program a robot to do human-level mathematical reasoning if you programmed it solely with algorithmic mathematical reasoning programs. But suppose that you programmed it solely with totally nonnormative brain simulator programs. There is no question of 'truth judgments' or 'soundness' in the programs. Nothing in his argument shows that 'human-level mathematical reasoning' could not emerge as a product of the robot's brain simulator program, just as it emerges as a product of actual human brains…From the fact that we cannot simulate a process at the level of, and under the description, 'theorem proving' or mathematical reasoning' it does not follow that we cannot simulate the very same process, with the same predictions, at some other level and under another description.[10]

Searle's comments strike me as more of a semantic diversion than a substantive philosophical objection. I am not at all opposed to utilizing thought experiments of varying degrees of plausibility in the course of one's reasoning – indeed, without

such experiments many of the most important developments in philosophy and the sciences would have been otherwise unattainable. Even so, it is not at all clear to me that the "nonnormative brain simulators" which Searle references above are the least bit conceivable. Other than denoting something antithetical to what we envisage as a "computer" today, what does this mean? Perhaps sensing the flimsiness of his rebuke of Penrose, Searle admits such a thought is akin to "science fiction fantasy." Fictional though they may be, thought experiments (at least the compelling ones) must bear some manner of *existential constituents*; we must be able to portray them in terms that go beyond a mere vacuous semantic husk. Where Penrose conveys the rough correlation "mind implies mathematical-reasoning aptitude"; Searle seems to me, by contrast, to effectively scream "nonnormative!" in a crowded theater of mathematicians and neuroscientists. Not only is his comment openly specious, it is, moreover, something of a philosophical non sequitur.

These qualms aside, Searle's criticism is also faulty on the level of analytic *depth*. He maintains that Penrose fails to prove that non-computational processes could not give rise to simulated, human-level mathematical reasoning. So then a counterexample to Penrose's thesis, which Searle entertains, would imply that human-level mathematical reasoning could exist within a system of mental states (or some process approximating mental states) that is *systemically* non-computational; in other words, such an embedding of processes are each or in total, devoid of any semblance of computational mathematics – at any *depth*. But then suppose, for the sake of argument, that we are once again sending an electrical

impulse through our trusty tube sock and that by application of binary logic, this process simulates integer addition. By Searle's lights, if we consider the simulation-system only at the level of "complexity" associated with the tube sock and "simpler" entities present, then, from the point of view of this level of system-depth, our simulation might appear "nonnormative." But it isn't. Searle wants us to ask questions along the lines of: does a tube sock *look* like 2+3 = 5? Do the fibers in the sock exert a *causal influence* upon this same mathematical equation? Obviously such considerations are absurd. The *system* itself, considered with respect to a higher degree of inclusion, is computationally rendered (*viz*., through the electrical signals interpreted via binary logic). Taken macroscopically, a simulation that bears absolutely no tokens of computational mathematics cannot engender human-level mathematical reasoning.

I would like at this juncture to further strengthen a rejection of weak A.I. from the perspective of the *converse* of our prior investigation of mind-computer equivalence. That is to say, I wish to show that in addition to the fact that the mind bears abilities that differentiate it from any computer simulation, there are abilities characteristic of computer simulations which are not reproducible by human minds. I propose three such examples.

The first example demonstrates an epistemological distinction between computers and minds. In 1976, amidst a storm of controversy, the *four color theorem* became the first major theorem of mathematics proven by computers. The theorem roughly states that given any separation of the plane into contiguous figures, no more than four colors are required to color such a *map* so that no

adjacent regions have the same color. The theorem was first proposed by Arthur Cayley in 1879, crediting August De Morgan, and two early attempted proofs were offered within a year of its inception. Although both proofs were initially accepted as valid within the mathematics community, flaws were nevertheless later discovered in each, and by 1890 both proofs were dismissed. For decades the theorem remained impervious to a formal proof, until at which point in the 1960s mathematicians began to develop methods involving computers. Apel and Haken later proved the theorem by reducing the set of all such possible configurations of the plane into maps down to 1,936 reducible constituent maps (meaning that any map in the plane can be reduced down to a union of this set of simple maps). Then, through a painstaking combinatorial process, computers confirmed that the theorem held for this particular set of maps; the result of the four color theorem then followed.

I want to be clear as to why this example proves that computers or Turing machines can achieve things that are qualitatively different from the types of things achievable by human minds. The fact that a Turing machine is more computationally efficient than you or I is not at all relevant to this purpose. Computers naturally have colossal (even potentially infinite) storage capacities and arithmetical potential. And yet, this feature does not represent an inherent structural difference from the way human minds perform similar mental operations; both processes involve an equivalent syntactic analysis. The point then is that in the case of the four color theorem, the computer possesses additional (present) *knowledge* beyond that of human minds (whether the four color theorem is unprovable by

human minds is another story – though it has been alleged that certain results, the Reimann hypothesis being the most notorious, are unprovable by human minds). Where the truth-value of the proposition: *the four color theorem is valid*, is undecided for human minds, possibly hopelessly so, such is not the case for Turing machines. Computers and human minds therefore possess different epistemic properties.

Consider now a second example. We have known of the irrationality of the number *pi* since Lambert's proof dating from 1761. Because the decimal expansion of an irrational number is non-recurring and interminable, we may conjecture (such a claim today remains however unproven) that this expansion approximates random number generation. If, for instance, one applies Kolmogorov's complexity criterion, it would follow that because it is impossible to reduce the informational content of the decimal expansion of *pi* to a proper subset of itself, this expansion is *infinitely complex*. It is not, for instance, known whether the decimal expansion of *pi* contains somewhere, say, one billion consecutive *1*'s. Despite the almost imponderably small probability that one could locate such a string of numbers, the existence of such a string is not a mathematical impossibility – so far as we know. Unassisted, a human mind will never discover, I would presume, the truth-value of the proposition: *the decimal expansion of pi contains one billion consecutive 1's* (call this property $\hat{p}$). On the other hand, it is conceivable that a computer – while it is impossible to disprove such a claim – could credibly confirm that $\hat{p}$ holds for *pi* (though not by using today's levels of processing power). A computer could, in other words, know something that we can't know – if in this instance, there is in

fact *something* to know. The main idea of course is that this potential difference is sufficient to indicate a qualitative difference between minds and computers.

Perhaps the reader will here allege that I have simply dressed up the prior example in something like a false nose and mustache; after all, haven't we already covered the case of epistemological differences between minds and computers? Yes and no. If we delve a little deeper, this example actually shows something different. Supposing that the computer finds such a string of *1*'s, it is then making at bottom an *existential claim* – a claim that is in no way computationally-contingent. Either the decimal expansion of *pi* has this property or it is does not, and this result exists independent of any computer. A devout Platonist might even be so bold as to allege that the result is mind-independent.

This situation is still thornier though. A basic property from order theory know as the *principle of trichotomy*, as applied to real numbers, states that all such numbers are either positive, negative or uniquely zero (for cardinals, trichotomy as an ordering principle is equivalent to the axiom of choice). Consider then the following "number" ϕ, defined as such:

$$\phi = \begin{cases} -1 & \text{if } \hat{p} \text{ holds for } \pi \\ 1 & \text{if } \hat{p} \text{ does not hold for } \pi \end{cases}$$

Note that from above, I claim that ϕ is *well-defined* in an existential sense since this definition is objectively unambiguous. Trichotomy says something structural about systems of numbers. That is to say, it discloses information about a formal system. But if you query a human mind as to whether ϕ satisfies the principle of

trichotomy for real numbers, the reply must be no, since ϕ is undetermined; thus for human minds trichotomy fails. Conversely, because a computer can *potentially* decide whether $\hat{p}$ is true of *pi*, a computer has no reason (equivalent examples notwithstanding) to "disbelieve" the principle of trichotomy. Accordingly, beginning from a set of comparable formal premises, a computer attains a structural asymmetry separating it from types of human mental processes. Ergo, computers and minds are different things.

For the third of the *converse* counterexamples of CTM, suppose, for the sake of argument, that the premises of the so-called Unified Field Theory are tenable. Then it would follow that the totality of the various states of *dark matter* in the universe could be known completely, and likewise for "luminous" or *light matter*. Such a UTE (unified theory of everything) would encapsulate the total cosmic equilibrium of dark and light matter; call this encapsulated equilibrium theory T_{ε}. From T_{ε} we devise a non-negative function, $C_{\varepsilon}(t)$, which cumulatively measures the "cosmic equilibrium rating" of the universe at time *t*, where $t = t_0$ corresponds with the instantaneous moment of the Big Bang. To better illustrate our example, presume further that this function is defined in such a way that if $\int_{t_0}^{t} C_{\varepsilon}(t)d\mu$ *converges* for some time *t*, where *t* is some temporal event that succeeds the Big Bang, then the proportion of dark matter to light matter in the universe at time *t* is *less than one* (where *one* indicates a totality of dark matter and *zero* a totality of light matter). Suppose also that if $\int_{t_0}^{t} C_{\varepsilon}(t)d\mu$ *diverges*, then the

proportion of dark matter to light matter in the universe at time *t* is *equal to one*. Due to the overwhelming complexity of such a function $C_{\varepsilon}(t)$, the value of $\int_{t_0}^{t} C_{\varepsilon}(t) d\mu$ for any arbitrary *t* would be indeterminable by any human mind and yet it would be determinable by a computer, if we allow for something approximating infinite computational potential. But then if this integral diverges for some *future state*, call it t_f, this serves to indicate that the universe would cease to exist at time t_f! (Note that if we take $\inf\{t_f\}$ over all such values, then the precise moment of quietus may be knowable, though this directly depends upon the nature of $C_{\varepsilon}(t)$ and many other factors). So it then follows that a computer *could* have specific knowledge of the ontological condition of the universe (namely, its total destruction) at some arbitrary future state, whereas no human mind could conceivably have access to this kind of ontological-knowing – unless of course, it asked a computer. Hence computers are not minds.

In closing, it is worth taking stock of several of the benefits and also some of the drawbacks and deficiencies of the arguments contained in the present paper. In the first place, the arguments in this paper rely on several major philosophy postures that are individually contestable. These attitudes include a commitment to some form of dualism, a *prima facie* acceptance of the preeminence of mathematics as a "mark of the mental," as well as a belief in the rectitude of the Turing test.

The rejection of weak, and not merely strong A.I., is a powerful, albeit somewhat controversial result. If one finds this argumentation compelling, then the endless morass of rejoinders to the dismissal of strong A.I. as instantiated, say, by

the Chinese Room example, become inconsequential.

The refutation of CTM has the additional advantage that it is attuned to many historical conceptions of the self – with the *cogito* being perhaps the most prominent of such formulations. By contrast, CTM bears the mark of the ahistoricity that is common to modern discourse; it is, in this way, part and parcel of a drive to view modernity, viz. the modern-self, as somehow removed from the "grand narrative" of history.

Postmodernism commonly annunciates the dawn of the epoch of the "expired subject," of the "vanishing author" and similar such happenings. Many of these approaches conceive of the *enigma of the self* as a phantom emerging from an complex of illusory ideological abstractions. This propensity is borne out through an effort to undermine the so-called "transcendental pretense of humanism." However, such inclinations frequently support a lateral movement toward the effacement of the self. I consequently wonder, together with Nietzsche: "if we never seek out a self – how could it happen that we should ever *find* ourselves?"[11] Our deeply-felt sense of self-awareness, of the unshakable "beingness of being" (Heidegger) challenges, for instance, the assertion that we are mere brains in a vat. And likewise it cannot be said that we are vacant computer simulations – as I believe some of the previous arguments bear out.

Notes

1. See Gregory Chaitin's *Meta Math: The Quest for Omega*.
2. Due to its more elaborate philosophical commitments, the "strong" version of CTM is the more commonly – and more easily – contested version.
3. Perhaps this oversight is due in part to the sheer quantity and multifariousness of the discoveries Turing made in the sciences.
4. See for instance Jack Copeland and Diane Proudfoot's article "What Turing Did After He Invented the Universal Turing Machine," or David King's "Is the Human Mind a Turing Machine?"
5. This example is duplicated from Heffernan, "Some Doubts About 'Turing Machine Arguments,' " p. 642.
6. John Searle, "Minds, Brains and Programs." *Philosophy of Mind: A Guide and Anthology*, p. 235.
7. Searle, p. 238.
8. This proof comes from Searle's article "Roger Penrose, Kurt Gödel, and the Cytoskeletons."
9. Searle, article "Roger Penrose, Kurt Gödel, and the Cytoskeletons," p. 74.
10. Searle, p. 80.
11. "We have never sought ourselves--how could it happen that we should ever find ourselves?" – Nietzsche, *On the Genealogy of Morality*, Prologue.
12. This version of Cantor's proof comes from Kaplasnky, *Set Theory and Metric Spaces*, p. 55.
13. Vladimir Tasic, *Mathematics and the Roots of Postmodern Thought*, p. 55
14. Tasic, p. 54

Bibliography

Block, Ned. "The Mind as Software in the Brain." *Philosophy of Mind: A Guide and Anthology*. Ed. John Heil. Oxford: Oxford University Press, 2004. 267-271

Boden, Margaret. "Escaping from the Chinese Room." *Philosophy of Mind: A Guide and Anthology*. Ed. John Heil. Oxford: Oxford University Press, 2004. 253-266

Chaitin, Gregory. *Meta Math: The Quest for Omega*. New York: Pantheon Books, 2005.

Clarke, J.J. "Turing Machines and the Mind-Body Problem." *The British Journal for the Philosophy of Science*, Vol. 23, No. 1, (Feb. 1972), pp. 1-12

Copeland, B. Jack and Diane Proudfoot. "What Turing Did After He Invented the Universal Turing Machine." *Journal of Logic, Language and Information*, Vol. 9, No. 4, Special Issue on Alan Turing and Artificial Intelligence (Oct. 2000), pp. 491-509

Feser, Edward. *Philosophy of Mind*. Oxford: Oneworld Publications, 2008.

George, Alexander and Daniel Velleman. *Philosophies of Mathematics*. New York: Wiley-Blackwell, 2001.

Harnad, S. "Minds, Machines and Turing: The Indistinguishability of Indistinguishables." *Journal of Logic, Language and Information*, Vol. 9, No. 4, Special Issue on Alan Turing and Artificial Intelligence (Oct. 2000), pp. 425-445

Heffernan, James. "Some Doubts About 'Turing Machine Arguments.' " Philosophy of Science, Vol. 45, No. 4 (Dec., 1978), pp. 638-647

Kaplansky, Irving. *Set Theory and Metric Spaces*. Providence: American Mathematical Society, 2001.

Kim, Jaegwon. *Philosophy of Mind*. Boulder: Westview Press, 2008.

King, David. "Is the Human Mind a Turing Machine?" *Synthese*, Vol. 208, No. 3, Computation, Cognition and AI (Sep, 1996), pp. 379-389

Kleene, Stephen Cole. *Introduction to MetaMathematics*. New York: Ishi Press International, 2009.

Lowe, E.J. *An Introduction to the Philosophy of Mind*. Cambridge: Cambridge University Press, 2000.

Nagel, Ernest and James Newman. *Gödel's Proof*. New York: NYU Press, 2008.

Ravenscrof, Ian. *Philosophy of Mind: A Beginner's Guide*. Oxford: Oxford University Press, 2005.

Searle, John. "Minds, Brains and Programs." *Philosophy of Mind: A Guide and Anthology*. Ed. John Heil. Oxford: Oxford University Press, 2004. 235-252

—-. *Mind: A Brief Introduction*. Oxford: Oxford University Press, 2004.

—-. "Roger Penrose, Kurt Gödel, and the Cytoskeletons." The Mystery of Consciousness. New York: The New York Review of Books, 1997. 55-93

Slezak, Peter. "Gödel's Theorem and the Mind." *The British Journal for the Philosophy of Science*, Vol. 33, No.1 (Mar., 1982), pp. 41-52

Tasic, Vladimir. *Mathematics and the Roots of Postmodern Thought*. Oxford: Oxford University Press, 2001.

Turner, Alan. "Computing Machinery and Intelligence." *Philosophy of Mind: A Guide and Anthology*. Ed. John Heil. Oxford: Oxford University Press, 2004. 212-234