PHILOSOPHICAL PSYCHOLOGY, VOL. 14, NO. 2, 2001

Are we evolved computers?: A critical review of Steven Pinker's *How the mind works*

SELMER BRINGSJORD

ABSTRACT *Steven Pinker's How the mind works (HTMW) marks in my opinion an historic point in the history of humankind's attempt to understand itself. Socrates delivered his "know thyself" imperative rather long ago, and now, finally, in this behemoth of a book, published at the dawn of a new millennium, Pinker steps up to have psychology tell us what we are: computers crafted by evolution—end of story; mystery solved; and the poor philosophers, having never managed to obey Socrates' command, are left alone to wander in the labyrinth of their benighted speculation forever. Unfortunately, though HTMW is to this point the crowning attempt of psychology to make systematic sense of persons by integrating everything relevant science knows, the book fails—and it fails so fundamentally and irremediably that we would do well to wonder anew whether we should supplant the basic view it promotes with what I call the super-mind hypothesis: the view that though mere animals are evolved computers, persons are more.*

1. Setting the stage

Steven Pinker's *How the mind works* (HTMW) seems to me to mark an historic point in the history of humankind's attempt understand itself. Socrates delivered his "know thyself" imperative rather long ago, and now, finally, in this behemoth of a book, published at the dawn of a new millennium, Pinker steps up to have psychology tell us what we are: computers crafted by evolution—end of story; mystery solved; and the poor philosophers, having never managed to obey Socrates' command, are left alone to wander in the labyrinth of their benighted speculation forever. Unfortunately, though *HTMW* is to this point the crowning attempt of psychology to make systematic sense of persons by integrating everything relevant science knows, the book fails—and it fails so fundamentally and irremediably that we would do well to wonder whether we should supplant the basic view it promotes with what I call the super-mind hypothesis: the view that though mere animals are evolved computers, persons are more.

Selmer Bringsjord, The Minds & Machines Laboratory, Department of Philosophy, Psychology & Cognitive Science, Department of Computer Science, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180-3590, USA, e-mail: selmer@rpi.edu.http://www.rpi.edu/~ brings

ISSN 0951-5089/print/ISSN 1465-394X/online/01/02020227-17 © 2001 Taylor & Francis Ltd
DOI: 10.1080/09515080120051580

TABLE 1. Some Bringsjordian deductive arguments against $\mathcal{C}$

Argument	Location
The zombie attack on $\mathcal{C}$	Bringsjord, 1999
The argument against $\mathcal{C}$ from free will	Chapter VIII, Bringsjord, 1992
Modified Chinese room argument against $\mathcal{C}$	Chapter V, Bringsjord, 1992
Argument against $\mathcal{C}$ From infinitary mathematical expertise	Bringsjord, 1997 *b*
The argument against $\mathcal{C}$ From irreversibility	Bringsjord & Zenzen, 1997

1.1. Pinker's position

Pinker unambiguously expresses his main position a number of times at the beginning of his book. For example, he tells us in the preface that the "the mind is a system of organs of computation designed by natural selection to solve problems faced by our evolutionary ancestors in their foraging way of life" (1997, p. x). And again soon thereafter in Chapter 1 we read:

> The mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life, in particular, understanding and ontmaneuvering objects, animals, plants, and other people. (1997, p. 21)

Pinker's main thesis is thus computationalism with an evolutionary twist. Let's label this thesis $\mathcal{C}^{\mathcal{E}}$, and let's use $\mathcal{C}$ to denote "straight" computationalism, the view that minds are computing machines, with $\mathcal{E}$ abbreviating the proposition that minds arose from evolution by mutation and natural selection. So, $\mathcal{C}^{\mathcal{E}}$ ($= \mathcal{C} \wedge \mathcal{E}$) entails $\mathcal{C}$ ($\mathcal{E}$), and hence by *modus tollens* any argument that overthrows $\mathcal{C}$ ($\mathcal{E}$) also overthrows $\mathcal{C}^{\mathcal{E}}$[1]. Now, as nearly all readers of this paper and *HTMW* know, many do reject $\mathcal{C}$ ($\mathcal{E}$)—and on the basis of *reasoned argument*, no less. For example, I reject $\mathcal{C}$ on the strength of, at last count, 15 rather carefully articulated deductive arguments, some of which are shown in Table 1. I also reject $\mathcal{E}$, and here again I'm far from alone. Soulmates include Kurt Gödel, Noam Chomsky, Roger Penrose, Alvin Plantinga, Paul Davies, Edgar Allen Poe, and many others—all of whom base their rejection not on some wild-eyed emotionalism or religious fanaticism, but on cool ratiocination [2].

1.2. Pinker's polemical strategy

And yet, despite this well-known fact (i.e. that there are countless careful, published attacks on $\mathcal{C}$ and $\mathcal{E}$ from first-rate minds), *HTMW* contains no explicit, sustained counterattacks, and no explicit arguments for $\mathcal{C}$ or $\mathcal{E}$. I say *sustained* counterattacks, because Pinker does devote a microscopic four pages to Searle's (1980) famous—indeed, possibly immortal[3]—Chinese room argument against $\mathcal{C}$, and he tosses out a quark-sized treatment (one page) of Penrose's (1994, 1989) well-known Gödelian attacks on this thesis. By contrast, Dennett's (1995) *Darwin's dangerous idea* (*DDI*), another book devoted to a defense of $\mathcal{C}^{\mathcal{E}}$ (but one not nearly as well-written, interesting, comprehensive, and powerful as *HTMW*), contains a full chapter devoted to a refutation of Penrose's Gödelian arguments. What's going on? What's going on is that Pinker is placing his chips on a different strategy; he's not trying to win at open, explicit, declarative debate. His strategy is to argue for $\mathcal{C}^{\mathcal{E}}$ by showing that it can work explanatory wonders. If we embrace $\mathcal{C}^{\mathcal{E}}$, then we can explain reason and emotion, art (from the truly sublime to mass entertainment), religion, imagery, marriage and adultery, rebellious children, deranged snipers, flights of creativity ... well, you get the idea: the list goes on.

1.3. The structure of the book: superficial vs. underlying

Reflection upon Pinker's position and polemical strategy reveals the underlying structure of his book. I distinguish between the *superficial* (or official) structure and the *underlying* structure. In the superficial structure, the chapters in his book, one by one, focus on explaining the computational view and showing all the explanatory fruit it can (supposedly) bear, but since the view can't be insulated from objection, I'll resist commentary on his facile applications of evolutionary psychology [4]. In the real, underlying structure, the foundation for Pinker's project is Chapter 5, in which Pinker presents and attempts to solve Wallace's paradox (WP), which Pinker (rightly) regards to be a direct and singular threat to $\mathcal{C}^{\mathcal{E}}$. This paradox takes its name from the co-inventor, with Darwin, of the theory of evolution: Alfred Russell Wallace[5]. WP arises from the existence of mental "overkill": mental powers which seem to have no explanation from the standpoint of evolution. Specifically, these powers seem to be massively superfluous for problems faced by our evolutionary ancestors in their foraging way of life. (Such foragers were called "savages" in Wallace's time.) Wallace pointed toward his paradox when he wrote:

> Our law, our government, and our science continually require us to reason through a variety of complicated phenomena to the expected result. Even our games, such as chess, compel us to exercise all these faculties in a remarkable degree. Compare this with the savage languages, which contain no words for abstract conceptions; the utter want of foresight of the savage man beyond his simplest necessities; his inability to combine, or to compare, or to reason on any general subject that does not immediately appeal to his senses ...
>
> ... A Brain one-half larger than that of the gorilla would ... fully have sufficed for the limited mental development of the savage; and we must therefore admit that the large brain he actually possesses could never have been solely developed by any of those laws of evolution, whose essence is, that they lead to a degree of organization exactly proportionate to the wants of each species, never beyond those wants ... Natural selection could only have endowed savage man with a brain a few degrees superior to that of an ape, whereas he actually possesses one very little inferior to that of a philosopher. (Wallace, in Pinker, 1997, p. 300) [6]

To his considerable credit, Pinker admits that WP is a fundamental threat to his enterprise. In fact, he calls WP a "central problem of psychology, biology, and the scientific worldview" (1997, p. 300), and a "foundation-shaking mystery" (1997)—but a mystery he promises to solve. As we'll see shortly, the promise isn't kept.

There is a *second* key to unlocking the underlying structure of *HTMW*. In the superficial structure, consideration of some mental phenomena are quietly postponed in Chapter 3 until the very end of the book—consciousness and free will, for example. If you measure the importance of such phenomena to Pinker's case for $C^{\AE}$ by how much space and energy he gives them, the result can be none other than that such phenomena are very nearly irrelevant. Nothing could be further from the truth, for two reasons. The first reason is simply that though Pinker may want to put consciousness and the like in the attic while he proceeds to polish up the part of the house that "matters" (the part that deals with perception, emotion, and so on), the brute fact of the matter is that no mental power or property is more important to us humans than consciousness. Indeed, the only reason any of us choose to go on living is to try see to it that we and others experience certain sorts of desirable conscious states! If you knew that starting tomorrow at 6 a.m. "you" were going to forever more have the consciousness of a current computing machine (the new SUN on my desk, e.g.), you would know that at 6:01 a.m. you would be *dead*—despite the fact that "you" *qua* computer would be able to manipulate all sorts of information[7]. So, the true structure of *HTMW* is one in which the last section of the book is not a poetic postscript, but rather a section of paramount importance. The second reason why the phenomena in question are crucial is that such things as qualia are intimately related to that which we've already observed to be at the core of Pinker's enterprise: Wallace's paradox. This is so because whatever justification there may be for regarding cognitive powers such as reasoning to be overkill from the standpoint of evolution would apply, *mutatis mutandis*, to such things as qualia. Why would our foraging ancestors need qualia to thrive? Why couldn't such creatures have had in their behavioral repertoire actions like finding berries and making fires and killing prey—all without qualia like the qualitative inner feelings associated with (say) sitting next to a fire? After all, when we build robots to search out counterparts to fire (power sources, say), do we worry about giving these robots associated inner *feelings*? These are of course rhetorical questions only (we will take them up in earnest below), but nonetheless the present point, I take it, is made: in the underlying structure of *HTMW*, the final section of the book looms very large indeed.

1.4. Plan for the remainder of the paper

The plan for the remainder of this paper is straightforward. In the next section, I show that Pinker doesn't solve WP in the least. In Section 3 I discuss Pinker's stance on philosophy and the "big questions" it traditionally treats, and explain why this stance, given the laws of elementary logic, vitiates his entire project. In the final section (4) I take stock of where my investigation leaves us.

2. Wallace's paradox still stands

2.1. Clarifying Wallace's paradox

Wallace's paradox, as it stands, is painfully obscure, and to my knowledge it hasn't been analyzed rigorously elsewhere[8]. I therefore begin this section with an attempt to get clearer about Wallace's paradox by way of a particular idealization. This idealization will appeal to *problems* faced by, and *powers* possessed by, persons and intelligent agents, information-processing creatures that form the central subject matter of the field of artificial intelligence (AI). Powers will be intimately connected to the notion of a *logical system*[9]. Suppose that problems start with the utterly easy and pass gradually on up to the tremendously hard in a continuum:

$$\mathcal{P} = p_1, p_2, p_3, \ldots .$$

At the easy end of the spectrum we might have a straightforward task of getting from point A to point B, or of detecting light and moving toward it, or perhaps the problems solved by simple intellgent agents in (Russell & Norvig 1994). Somewhere after problems like these we would find problems like those representable and solvable in the propositional calculus, and those involved in the real-world analogue of a foraging way of life. Moving toward the part of $\mathcal{P}$ populated by truly difficult problems, we would have problems requiring more powerful logical systems than the propositional calculus (e.g. first-order logic), and then we would head on up toward problems solved by novelists, professional logicians, mathematicians, and so on.

For an example of a problem requiring a logical system as powerful as first-order logic[10], I offer the "Dreadsbury Mansion Mystery," which would seem to at least to a slight degree resemble problems solved by "real-world" detectives[11]. Here is the mystery.

> Someone who lives in Dreadsbury Mansion killed Aunt Agatha. Agatha, the butler, and Charles live in Dreadsbury Mansion, and are the only people who live therein. A killer always hates his victim, and is never richer than his victim. Charles hates no one that Aunt Agatha hates. Agatha hates everyone except the butler. The butler hates everyone not richer than Aunt Agatha. The butler hates everyone Agatha hates. No one hates everyone. Agatha is not the butler.
>
> Now, given the above clues, there is a bit of a disagreement between three Norwegian detectives: Inspector Bjorn is sure that Charles didn't do it. Is he right? Inspector Reider is sure that it was a suicide. Is he right? Inspector Olaf is sure that the butler, despite conventional wisdom, is innocent. Is he right?

If you would like to try to solve this problem on your own, give it a shot now, and check your reasoning by looking at the footnote referenced at the end of this sentence[12].

And now what would be an example of a truly difficult problem, and its solution? I would suggest the problem Gödel solved by way of his famous first incompleteness theorem (= Gödel I). I have elsewhere explained in considerable technical detail why the reasoning in Gödel's solution can't even be approximated by today's automated theorem provers (Bringsjord, 1998). (Note that this result is quite different than the Penrose-alleged (1994) implication of $\neg\mathcal{C}$ from Gödel I. I've treated this implication in (Bringsjord & Xiao, 2000). Unfortunately, to specify this example here isn't practicable: there is just too much knowledge and expertise presupposed by the full presentation of such a thing, and too little space available here (and hence I refer motivated readers to any or all of Boolos & Jeffrey, 1989; Ebbinghaus, *et al.*, 1984; Smullyan 1992)).

Now let's return to the second spectrum we need, one for the powers side. Here we similarly invoke a continuum

$$O = o_1, o_2, o_3, \ldots$$

of powers, running from the simplest of powers, to more impressive ones (as in those possessed by robots able to reason in a fragment of the propositional calculus), to powers (e.g.) professional mathematicians and logicians exploit. If you've assimilated Gödel's incompleteness theorems, you probably have a sense for what powers appear at the upper end of the continuum. In order to solve even problems at the upper end of the continuum (= problems solved by professional logicians and mathematicians), one needs to move between, create, reason about (etc.) many different logical systems.

Next we invoke a function e that maps $\mathcal{P}$ to the power set of O (2^O). The idea here is that this function yields, for a given problem p_i, the stretch of contiguous o_i's that are powers that might possibly be "crafted" by evolution to solve p_i.

At this point we can set out Wallace's paradox in the form of the sort of deductive progression traditionally used in carefully setting out paradoxes:

(1) If people evolved, then for every power o_i possessed by our foraging ancestors, there exists a problem p_k faced by them or their predecessors such that $o_i \in e(p_k)$.

(2) It's not the case that for every power o_i possessed by our foraging ancestors, there exists a problem p_k faced by them or their predecessors such that $o_i \in e(p_k)$.

(3) $\therefore$ People didn't evolve. (from 1 and 2)

Of course, the idea is that this constitutes a *paradox* because there is supposed to be consensus that people evolved. That is, those who affirm $\mathcal{E}$ (or $\mathcal{C}^{\mathcal{E}}$) are faced with having to swallow a contradiction (while those who are skeptical about $\mathcal{E}$ see in WP a direct argument for $\neg\mathcal{E}$, and therefore for $\neg\mathcal{C}^{\mathcal{E}}$).

2.2. *Classifying reaction to Wallace's paradox*

Is Wallace's paradox sound? Well, there's no question that the argument is formally valid: the core inference is *modus tollens*; and it's clear that a formal version of the argument (represented in first-order logic, e.g.) would certify the quantifier reasoning used[13]. But what about the premises? Are they true? There would seem to be

TABLE 2. Reaction to wallace's paradox

View	Reductionism	Exaptationism	Exotic natural forces	Theism
Response to WP	Rejects (2)	Rejects (1)	Accepts (3)	Accepts (3)
Sample proponent	Pinker	Gould	Penrose, Chomsky	Plantinga, Davies

four main alternatives in the face of this question; they are shown in Table 2[14]. In two of the answers, exotic natural forces and theism, the premises are accepted; given that the argument is formally valid, this amounts to accepting (3), the conclusion. In each of the other two answers, one of the premises is rejected. Pinker himself rejects premise (2), under the view I call "reductionism." Gould rejects premise (1), under the view known as "exaptationism." According to reductionism, even the greatest mental powers we have can be reduced to powers used directly by our foraging ancestors in their concrete behaviors. According to exaptationism, while it's true that the powers possessed by our foraging ancestors far exceeded the problems that needed to be solved to survive in their environment, evolution fortuitously gave them powers that *we* can use to improve *our* lot.

These, then, are the four alternatives. Which one should a rational person affirm? Well, to start, Pinker himself rejects exaptationism. He writes:

> Stephen Jay Gould, in an illuminating essay on Darwin and Wallace, sees Wallace as an extreme adaptationist who ignores the possibility of exaptations: adaptive structures that are "fortuitously suited to other roles if elaborated" (such as jaw bones becoming middle-ear bones) and "features that arise without functions ... but remain available for later co-optation" (such as the panda's thumb, which is really a jury-rigged wristbone).
>
> Objects designed for definite purposes can, as a result of their structural complexity, perform many other tasks as well. A factory may install a computer only to issue the monthly pay checks, but such a machine can also analyze the election returns or whip anyone's ass (or at least perpetually tie them) in tic-tac-toe.
>
> I agree with Gould that the brain has been exapted for novelties like calculus or chess, but this is just an avowal of faith by people like us who believe in natural selection [= $\mathcal{E}$]; it can hardly fail to be true. It raises the issue of who or what is doing the elaborating and co-opting, and why the original structures were suited to being co-opted. The factory analogy is not helpful. A computer that issues paychecks *cannot* also analyze the election returns or play tic-tac-toe, unless someone has reprogrammed it first. (Pinker, 1997, p. 301)

Here I'm with Pinker, and I think we all should be: he encapsulates some devastating objections. It's easy to see that exaptationism is really, at bottom, a leap of self-serving faith. After all, the exaptationist can *always* reply in Gould-like fashion to any candidate for a counter-example to $\mathcal{E}$. Faced with an astonishing power $o^\star$

possessed by a creature C able to thereby solve a problem $p^\star$ now, where $p^\star$ has no precedent in C's ancestors, the exaptationist can *always* claim that that $o^\star$ was present but untapped in one or more of these ancestors, and that its presence there was due to good luck. By the lights of Exaptationism, then, $\mathcal{E}$ is unfalsifiable, which would take this view outside the domain of science and philosophy. As Pinker points out, part of the root problem infecting such claims is that they leave aside a meaningful explanation as to "who or what" is responsible for these fortuitous powers[15].

Pinker is also right that the factory analogy fails. This point against Gould can be broadened and sharpened: even a simple multi-cellular organism includes "biological computers" sufficient to allow for the running of any Turing-computable function, given an appropriate program[16]. After all, an abacus, with a suitable set of instructions for how to manipulate beads in it, can compute any Turing-computable function[17]. But again, where would the instructions come from? Who or what would be responsible?

2.3. The problems infecting reductionism

Now let's turn to Pinker's response to WP, which he gives on pages 358–360. His response, which we've labeled 'reductionism' (recall Table 2), amounts to the claim that in the minds of our foraging ancestors were the very same fundamental powers that (e.g.) professional logicians of today exploit. Whereas our ancestors used these powers to do such things as track and kill animals, today people like Gödel use such powers to solve problems of the sort he solved. Pinker writes:

> The answer to the question [that sums up WP] "Why is the human mind adapted to think about arbitrary abstract entities?" is that it really isn't. Unlike computers and the rules of mathematical logic, we don't think in *F*'s and *x*'s and *y*'s. We have inherited a pad of forms that capture the key features of encounters among objects and forces, and the features of other consequential themes of the human condition such as fighting, food, and health. By erasing the contents and filling in the blanks with new symbols, we can adapt our inherited forms to more abstruse domains. (Pinker, 1997, p. 358)

Unfortunately for Pinker, his Reductionism fares no better than exaptationism. In fact, the former is plagued by some of the same fatal problems as those that explode the latter. For example, take the issue of falsifiability. If reductionism is is to have any bite at all, it must amount to something that could, at least in principle, be overthrown. Suppose, once again, that we observe, today, an astonishing power $o^\star$ possessed by a creature C able to bring this power to bear in solving a problem $p^\star$, where $p^\star$ has no precedent in C's ancestors. Is it enough, given this observation, for the proponent of reductionism to say "Oh, no problem, for you see, $o^\star$ is really, at bottom, an ability to manipulate a pad of forms used in epochs gone by to solve much simpler problems." It can't be enough to simply *say* this kind of thing—because if so, Reductionism is no more falsifiable than exaptationism: one could say this about any $p^\star$ and $o^\star$. So, let's try to give Pinker a fighting chance by fleshing out Reductionism at least a bit. We can do so by appealing to the machinery we set out earlier.

What kind of powers, then, would be associated with this wondrous "pad of forms" Pinker alludes to? What kind of logical system would the slots be in? For example, the propositional calculus has propositional variables for slots, but these slots and this system aren't sufficient for syllogistic reasoning, which people (to a degree) do *naturally* do. For example, we know that the syllogism

$$\begin{array}{c} \text{All } A\text{s are } B\text{s} \\ \text{All } B\text{s are } C\text{s} \\ \hline \text{All } A\text{s are } C\text{s} \end{array}$$

is valid, but a "pad of forms" corresponding to the propositional calculus is inadequate to represent this syllogism. For to represent it with such a "pad" would be to plug the statements within it into propositional variables, but the schema that results, namely,

$$\begin{array}{c} p \\ q \\ \hline \mathrm{r}, \end{array}$$

is about as invalid as deductive reasoning can get.

So what sorts of "blanks" are we talking about? What sorts of blanks and forms and slots does Pinker have in mind? He doesn't tell us; he tells us absolutely nothing about these schemes. And this silence is quite astonishing. After all, given the underlying structure of *HTMW*, the key issue has now boiled down to whether or not Reductionism is a cogent response to Wallace's paradox. (Of course, it wouldn't help Pinker in the least if either of the remaining two responses is correct.) If it *isn't* a cogent response, not only does the book itself fail, but by Pinker's lights, *psychology* fails. Darwin, faced with WP (delivered to Darwin by Wallace himself), had confidently proclaimed that before long the paradox would be solved by psychology, as this field had now been placed by the theory of evolution on "a new foundation." If reductionism doesn't solve WP, Darwin looks like little more than a wishful thinker. So, can some set of "blanks" of the type Pinker seems to have in mind do the trick, and save Darwin? I don't think any such set could possibly allow for the kind of cognition at the heart of logic and mathematics. In the next section, I explain why.

2.3.1 Reductionism can't reduce infinitary reasoning. Let's take some pains to consider some particular abstruse reasoning, and whether or not it can be captured by filling in the kind of blanks Pinker points toward. Our example is discussed in (Bringsjord, 1997*b*); we have enough space here to just quickly describe the basic idea behind the relevant reasoning—but this quick description should demonstrate that reductionism, at best, is wholly implausible.

Pinker's reductionism, when charitably understood, can do no better than appeal to first-order logic as the fleshing out of this mysterious “pad of forms.” At best, Pinker seems to have this logic in mind when, in the quote above, he says “Unlike computers and the rules of mathematical logic, we don’t think in *F*’s and *x*’s and *y*’s” (1997, p. 358) Computers do operate at the level of first-order logic. In fact, mathematically represented, computation can be completely carried out in terms of first-order logic. But the “pad of forms” that must be used to explain the behavior of logicians and mathematicians who prove things in and about *infinitary* logics is provably beyond first-order logic. One such logic is $\mathcal{L}_{w_1 w}$.

The basic idea behind $\mathcal{L}_{w_1 w}$ is straightforward. This system allows for infinite disjunctions and conjunctions,[18] where these disjunctions and conjunctions are no longer than the size of the set of natural numbers (let’s use w to denote the size of the set of natural numbers)[19]. Here is one simple formula in $\mathcal{L}_{w_1 w}$ which is such that any world that models it is finite:

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

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 the infinitary formula above cannot be captured by a formula in a system at or below first-order logic. 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 $\mathcal{L}_{w_1 w}$), reductionism is doomed (again, for more, see Bringsjord, 1997*b*; Bringsjord & Zenzen, 2001).

2.3.2. The second problem plaguing reductionism. Even if we suppose that by some miracle Pinker’s Reductionism can accommodate such powers as those tapped by logicians and mathematicians reasoning in and about infinitary logics, that such reasoning can be reduced to Pinker’s “slot system,” his response to WP still fails. The reason is that this slot system, whatever it is, didn’t need to be such that it *could* be extended to handle such things as infinitary logic. Even if one supposes that our use of (say) infinitary logic is somehow an inscrutable modification of first-order logic, there is a problem facing Pinker (and Darwin): What survival value did this modifiability have at the time our foraging ancestors were running about? Our foraging ancestors could have done just dandy with a slot system frozen at the level of expressiveness required for finding berries and killing animals. These foragers could have been like the robots in my lab: they are able to succeed in simple environments by way of inflexible powers *specifically engineered* for such environments. Pinker’s problem here is once again essentially the same as one he showed to be infecting Gould’s exaptationism. Recall Gould’s defective factory analogy, intended by him to support the notion that as the computers driving a payroll system could be harnessed to analyze election returns, evolution could have given our foraging ancestors powers for tracking animals that could later be used for the tensor calculus. Just as here Gould fails to understand the profound differences between programs for payroll systems versus those for election analysis, Pinker fails to understand the vast differences between programs dedicated to securing success in the foraging game *but not (say) in the tensor calculus game*, versus general-purpose programs able to secure success in *both* games.

The point here would seem to be one that AI and robotics makes even tougher for Pinker. I say this because progress in these fields would seem to be fairly interpreted as pressing against Pinker the disturbing possibility that we will soon be able to create a robot able to hunt and gather and communicate just as well as early *Homo sapiens*—but these robots will utterly lack the capacity to do mathematics of a sort suggested by Gödel's results, and demanded by open questions regarding $\mathcal{L}_{w_1 w}$. Why weren't our ancestors like these robots? Why did they need this strange and wondrous capacity for such things as Gödelian results? These are questions Pinker doesn't, and probably can't, answer.

3. Big questions

Pinker identifies psychology with what he calls "reverse engineering"[20]:

> On [the $\mathcal{C}^{\mathcal{L}}$] view, psychology is engineering in reverse. In forward-engineering, one designs a machine to do something; in reverse-engineering, one figures out what a machine was designed to do. Reverse-engineering is what the boffins at Sony do when a new product is announced by Panasonic, or vice versa. They buy one, bring it back to the lab, take screwdriver to it, and try to figure out what all the parts are for and how they combine to make the device work ... The rationale for reverse-engineering living things comes, of course, from Charles Darwin ... Darwin insisted that his theory explained not just the complexity of an animal's body but the complexity of its mind. "Psychology will be based on a new foundation," he famously predicted at the end of *The Origin of Species*. (Pinker, 1997, pp. 21–22)

What Pinker somehow misses is that if this is what psychology is, the field is failing miserably. This is easy to see. If one reverse engineers X, as Pinker has reminded us, one dissects X into parts X_1 X_2, ..., X_n, and at the same time, by drawing upon a library of primitive parts and concepts, one shows how X_i can be built from this library. *But we aren't able to pull off any such thing for an assignment of* X *to free will, consciousness, or abstract reasoning at the hard end of the continuum of powers we conceived earlier.*

Ironically, confirmation of this inability can be found by taking up a challenge Pinker himself sets in *HTMW*: his "Robot Challenge." This challenge is to build a robot with all the powers and abilities of persons. But three distinctive powers of people are the three just enumerated—and no AInik or roboticist has any idea how to engineer a system that has one of more of these powers!

What would be Pinker's response to this threat to his project? Fortunately, we don't need to speculate on his behalf; the answer is obvious. Pinker would say that though the reverse-engineering can be carried out *in principle*, it cannot be carried out in practice because of the doctrine of "cognitive closure." This is the doctrine that certain problems are simply too hard for people to *ever* solve. We read:

> Maybe philosophical problems are hard not because they are divine or irreducible or meaningless or workaday science, but because the mind of *Homo sapiens* lacks the cognitive equipment to solve them. We are organisms, not angels, and our minds are organs, not pipelines to the truth. Our minds evolved by natural selection to solve problems that were life-and-death matters to our ancestors, not to commune with correctness or to answer any question we are capable of asking. We cannot hold ten thousand words in short-term memory. We cannot see in ultraviolet light. We cannot mentally rotate an object in the fourth dimension. And perhaps we cannot solve conundrums like free will and sentience. (Pinker, 1997, p. 561)

It's exceedingly hard to see how this gets Pinker anywhere. He faces an acute problem: the overall thesis of his book, $\mathcal{C}^E$, that persons are evolved computers, is to be defended by showing how personhood can be reverse-engineered into components appealing only to the rudiments of evolution and computation. But those properties that are distinctive of personhood cannot be reverse-engineered. It does no good to say that if $\mathcal{C}^E$ is true, these properties would in fact turn out to be recalcitrant from the standpoint of reverse-engineering. Suppose I maintain that people have the remarkable abilities they do because super-fast microscopic gremlins are busy working in our brains. And suppose that my chief argument for this startling proposition is that the behavior of people can be dissected into gremlin behavior (for which I have descriptions). What progress do I make if, when faced with the fact that I cannot carry out the dissection for a host of human powers, I declare: "Well, what do you expect! Our thinking is really just a bunch of gremlins at work, and gremlin behavior is just not up to the task of reverse-engineering these powers!" The fallacy is patent. Students who reason in such ways fail Philosophy 101.

In the end, then, Pinker writes off philosophy as a monumental waste of time. He says that "no progress" has been made on the "big questions"—and that no progress *will* be made, given cognitive closure. The big questions are to be left aside as eternally unsolvable mysteries for philosophers to ponder unproductively for perpetuity. Now I myself am unsure about the merits of what Pinker calls philosophy. But I'm I'll-stake-my-life-sure about the efficacy and primacy of a *part* of philosophy: logic. I think we're smart enough to deduce that we're not evolved computers (recall Table 1), whereas Pinker believes that since we're evolved computers we're not smart enough to figure out whether or not we're evolved computers! It's the fundamental difference between confidence in armchair deduction, and confidence in empirical tinkering and induction. I think the deduction destroys $\mathcal{C}^E$. Pinker, as we've noted, not only affirms this proposition, but offers an argument of his own. Unfortunately for Pinker, logic trumps all; and logic shows that his argument is fallacious, *even if by some miracle it could be shown that evolution and computation explain personhood*; here's why. His overall argument is that $\mathcal{C}^{E}$ explains personhood (where the explanation comes in the form of blueprints for reverse-engineering personhood from the twin pillars of evolution and computation). But we know from logic that from that fact that p explains q one cannot infer that p is true. This leaves Pinker in all sorts of hot water. For example, suppose, with Pinker, that the God of Christian–Judaic–Islamic monotheism doesn't exist. (Pinker assumes that this God doesn't exist, and then of course tries to explain why so many humans believe He does.) Nonetheless, if we embrace the proposition that this God exists, we can explain a lot of things (e.g. human beings would exist because God desired to commune with them). Ergo, by Pinker's own reasoning we arrive at the contradiction that God both does and doesn't exist.

The general point here, that Pinker conflates truth and explanation, can be expressed in a rather different way. Imagine a rather nasty logician. This fellow is a fiendish egoist: he does what he wants to do, even if it means other people must die, or be tortured, and so on. If we confront this person with Pinker's book, and have a discussion, we will make no progress. Suppose the nasty logician asks, when we insist that he should not, say, murder to get ahead in the business world: "Why should I be moral?" Pinker may succeed in explaining to this unsavory chap why, in general, he behaves the way he does. But even if evolutionary explanations for morality are right, even if explanations along the line that morality has a special status as a system that maximizes the probability that *Homo sapiens* will grow and prosper on the whole are legitimate, how do we answer the brilliant egoistic fiend? Why should he listen if he can get away with his murders? Pinker can tell him how it is that there is a code of behavior that this fiend is casting aside, but Pinker must of necessity be silent when the fiendish logicians asks, smiling, why is it that *he* should abide by this code. Here again there is no answer to be found in an explanation (whether or not it appeals evolution and computation) rooted in the empirical. The answer can come only from a truth-seeking enterprise conducted from the armchair.

4. Persons as super-minds

Let's conclude by taking stock of where we've ended up.

First, return to the central disjunction of Table 2. If reductionism and exaptationism are now out of the picture, and if we're unwilling to accept a contradiction, dodging Wallace's paradox can be accomplished only if we either invoke Theism or exotic natural forces (or some synthesis of the two). We end up in this extraordinary position because personhood is beyond the reach of evolution and computation, as shown by our modernized version of Wallace's paradox.

Why do the duo of evolution and computation apparently work for reverse-engineering a rat or a parrot or a chimp, but fail when we're talking about people? The answer revealed by our analysis of *HTMW* is presumably that evolution and computation are explanatorily impotent in the face of the properties that distinguish persons. What are these properties? There are many, but herein we have focused upon only three within the list given (e.g.) in (Bringsjord, 1997 *a*), three which are directly relevant to *HTMW*, namely[21],

1. free will;
2. phenomenal consciousness; and
3. robust abstract reasoning (e.g. ability to create logical systems beyond first-order logic, and to switch between them on a principled basis).

It's exceedingly hard to see how these three properties can be captured by evolution and computation. Computers, whether evolved or not, don't seem to originate anything; they seem to do just what their programs instruct them to do[22]. So it's hard to see how they can originate decisions and actions ("free will"). It's also hard to see how phenomenal consciousness can be captured in any third-person scheme whatever, let alone in something as austere as computation[23]. And those in AI who seek to model abstract reasoning know well that we haven't even begun to show how sophisticated abstract reasoning like that involved in dealing with infinitary logics like $L_{w_1 w}$ can be cast in well-understood computable logics. Given what we know at present, some of this reasoning seems to be beyond the reach of computation[24]. Certainly such reasoning cannot be cashed out through Pinker's naive "pad of forms." This result implies that we ought to consider what I call the "super-mind hypothesis," according to which persons have information-processing powers beyond the reach of computation (e.g. the power to reason in and about infinitary logics), as well as powers (e.g. phenomenal consciousness) that cannot be expressed as information processing of any sort[25]. In the end, psychology hasn't the tools to reverse-engineer people, and Darwin, despite Pinker's mammoth tome, should be seen as someone who grasped the mechanical nature of animals, but expressed mere groundless faith that persons were at bottom the same.

Acknowledgements

I'm indebted to students in two of my *HTMW*-based seminars, and to others with whom I've discussed and debated this book: the remarkable summer 2000 researchers in my Minds & Machines Laboratory (Paul Bello, Paul Boyle, Devin Croak, Michael Dawe; Marc Dastefano, Ken Eitelman, Allyn Hill, Greg Oberg, Tom Poltrino, Kelsey Rinella, David Strong), Michael Zenzen, Jim Fahey, Jesse Prinz, and Steven Pinker himself, at whose question—"Well, but what else could the mind *be* than a computer?"—the super-mind hypothesis is aimed. The full version of the present paper, roughly twice the length of this published version, is available from the author and can currently be found at http://www.rpi.edu/~faheyj2/SB/SELPAP/PINKER/pinker.rev2.pdf.

Notes

[1] Though such an issue is outside the scope of the present paper, it would also seem to be the case that¬$\mathcal{C}$ implies¬$\mathcal{E}$. Why? Because evolution is itself a computable process, and no computable

process can produce a device capable of information-processing feats going beyond the computable.

[2] I don't have sufficient space to discuss the ratiocination in question. Let me say only that both Pinker's *HTMW* and Dennett's (1995) like-minded *Darwin's dangerous idea* (*DDI*) mark in some real sense the synthesis of Darwin and Turing, and perhaps the combined genius of this august pair is so great as to bestow upon these books a *prima facie* plausibility. But such an appeal to authority, over and above the fact that it constitutes a fallacy, would be dangerous for proponents of *HTMW* and *DDI* (and, thereby, $\mathcal{C}^{\mathcal{E}}$): Gödel was as great a mind as any that has graced our planet, and certainly stands as the terrestrial god of demonstration, yet he regarded both $\mathcal{C}$ and $\mathcal{E}$ to be demonstrably false for mathematical reasons (see Wang, 1974, p. 326).

[3] Soon to appear, more than two decades after its debut, from Oxford University Press, is a book devoted to Searle's CRA, edited by Mark Bishop and John Preston (forthcoming).

[4] By the way, there is no chapter on language. This might strike some of the same readers as odd, given, for example, that it's the linguistic powers of people that impel the likes of Noam Chomsky to toy with rejecting $\mathcal{E}$ (Chomsky's attitude is discussed in Dennett, 1995). The reason *HTMW* doesn't cover such powers is that Pinker sees himself as having treated them in a previous book (Pinker, 1994).

[5] Darwin received a rather shocking letter from Wallace in which was set out the essentials of evolution through mutation and natural selection—a letter which prompted the delaying Darwin to quickly hammer out *Origin* (1859)—and now one and a half centuries later, few and far between are the schoolchildren (or, for that matter, members of the intelligentsia) who know of Wallace. The Darwin–Wallace relationship was a rather complicated one; for two accounts, see Desmond and Moore (1991) and (Richards, 1987).

[6] That Wallace links the full power of the brains of *Homo sapiens* to the cognitive activities of philosophy is, I dare say, deliciously ironic given the roles Pinker prescribes for psychology and philosophy in *HTMW*.

[7] For a full treatment of the relationship between such gedanken-experiments and $\mathcal{C}$, see Bringsjord (1999).

[8] My own formal analysis can be found in the full version of the present paper, available directly from me and currently at at http://www.rpi.edu/ ~ faheyj2/SB/SELPAP/PINKER/pinker.rev2.pdf. In addition, the version intimated in the quote above from Wallace seems to contain a premise (that the "laws of evolution ... lead to a degree of organization exactly proportionate to the wants of each species, *never beyond those wants* ..."; emphasis mine to highlight the doubtful part of this premise) that is false. This premise isn't needed in the more sophisticated version of WP that I articulate below.

[9] Some logical systems are much in use in psychology (esp. of reasoning) and philosophy, and and will be familiar to many readers. For example, there is the propositional calculus, which allows for simple declarative statements to be represented by propositional variables p_1, p_2, ... which can be put together with help from the familiar truth-functional connectives $\neg$ (not), $\vee$ (or), $\wedge$ (and), $\rightarrow$ (if then), $\leftrightarrow$ (if and only if). (Of course, the form of the propositional variables can vary in accordance with desired expressivity.) Various methods for constructing proofs in the propositional calculus are possible; all methods invoke rules in one way or another. One foundational inference rule says that from a pair of formulas having the form $\phi \rightarrow \psi$ and ϕ one can infer ψ; this rule is known as *modus ponens*. While *modus ponens* makes no use of a supposition, one rule which does is known as *conditional proof*, according to which, if, after supposing ϕ, one can deduce ψ, then one can "discharge" the supposition and infer $\phi \rightarrow \psi$.

[10] As was the case with respect to the propositional calculus, I assume my readers to be to some degree familiar with first-order, or ordinary quantifier, logic, which adds to the machinery of the propositional calculus the quantifiers $\forall x$ (for all x) and $\exists x$ (there exists an x), and also predicate letters to represent individual properties. With these additions, statements in natural language impossible to capture in the propositional calculus (e.g. "All students are bored by at least, two professors") can be captured in first-order logic (e.g. the the example in the previous parenthetical: $\forall x\ (Sx \rightarrow (\ \exists y \exists z (y \neq z \wedge (Py \wedge Pz \wedge Byx \wedge Bzx))))$.

[11] I've "Norwegianized" the problem as it appears in (Pelletier, 1986). For a fuller discussion of such real-world reasoning in connection with Al and logical systems see (Bringsjord & Ferrucci, 2000).

[12] Here is a solution in the form of an informal proof that appeals to rules of inference in first-order logic. Suppose Charles is the killer. Then, given that a killer always hates his victim, Charles hates Agatha. We know that Agatha hates herself (because she hates everyone save for the butler). But then we have a contradiction: Charles hates no one that Aunt Agatha hates (given), and yet Charles, we've just deduced, hates Agatha, as does she. Ergo, by *reductioad absurdum*, Charles didn't kill Agatha. Suppose, on the other hand, that the butler killed Agatha. Then the butler hates Agatha, and is not richer than Agatha. Since we are given that the butler hates everyone not richer than Agatha, the butler hates himself (by universal quantifier elimination). The butler also hates Charles, because Agatha hates Charles. But then someone (the butler) hates everyone—a contradiction. Therefore, the butler didn't kill Agatha. Hence (by disjunctive syllogism) Agatha did kill herself; and so all the detectives are correct. This informal proof can easily be specified in the form of a formal proof in first-order logic, given obvious symbolization; e.g. for starters, "No one hates everyone" could be $\neg\exists x(Hxa \wedge Hxb \wedge Hxc)$.

[13] At least potentially, this may place Pinker in a rather ironic situation. As our analysis has just disclosed, in order to grapple with the issues raised by Pinker's book (i.e. WP!), one must use powers high on the spectrum of powers we've invoked. It would seem, specifically, that these powers must make use of parts of a logical systems at and above first-order logic—and yet Pinker seems to limit our foraging ancestors to reasoning in first-order logic and below.

[14] An excellent discussion of the exotic natural forces and theism views, replete with references and quotes (from, in some cases, unpublished conversations) involving Penrose, Chomsky, and Davies can be found in Demmett (1995). Plantinga's attack on $\mathcal{E}$ can be found in (Plantinga 1993).

[15] Jim Fahey has pointed out to me that such condemnations of exaptationism probably presuppose the Principle of Sufficient Reason. Though I don't believe this is the case, I happen to affirm certain versions of PSR.

[16] The class of functions that can be computed by any computer is traditionally identified with the class of functions that a Turing machine can compute (the Turing-computable functions). A Turing machine is a stunningly simple device that reads and writes symbols on a tape in accordance with austere instructions. The scheme was introduced by Turing (1936). A nice formal presentation of Turing machines can be found in many books (one good one is Lewis & Papadimitriou, 1981).

[17] The proof can be found in Boolos and Jeffrey (1989).

[18] Of course, even finitary logics have underlying alphabets that are infinite in size (the propositional calculus comes with an infinite supply of propositional variables). $L_{w_1 w}$, however, allows for formulas of infinite length—and hence allows for infinitely long derivations.

[19] This isn't the place to baptize some readers into the world of cardinal numbers. Hence we leave the size implications of the subscripts in $\mathcal{L}_{w_1 w}$, and other related niceties, such as the precise meaning of w therein, to the side. For a comprehensive array of the possibilities arising from varying the subscripts, see Dickmann (1975).

[20] Daniel Dennett and Marvin Minsky, among others, see psychology similarly. See Dennett (1995, pp. 386–387).

[21] The account is streamlined in the interests of space. For example, because people sleep (and because they can be hypnotized, etc.), a person would be a creature with the *capacity* to have properties like those listed here.

[22] For more on this, see Bringsjord (forthcoming), and also the second entry in Table 1.

[23] For more, see the bottom entry in Table 1.

[24] For more, see the fourth entry in Table 1.

[25] For specification and defense of the super-mind hypothesis, see Bringsjord and Zenzen (2001).

References

BISHOP, M. & PRESTON, J. (Eds) (forthcoming). *The Chinese room: new essays on John Searle's arguments against "Strong AI"*. Oxford: Oxford University Press.

BOOLOS, G.S. & JEFFREY, R.C. (1989). *Computability and logic.* Cambridge: Cambridge University Press.

BRINGSJORD, S. (1992). *What robots can and can't be.* Dordrecht: Kluwer.

BRINGSJORD, S. (1997 *a*). *Abortion: a dialogue*, Indianapolis, IN: Hackett.

BRINGSJORD, S. (1997*b*). An argument for the uncomputability of infinitary mathematical expertise. In P. FELTOVICH, K. FORD & P. HAYES (Eds) *Expertise in context.* Menlo Park, CA: AAAI Press.

BRINGSJORD, S. (1998). Is Gödelian model-based deductive reasoning computational? *Philosophica* **61**, 51–76.

BRINGSJORD, S. (1999). The zombie attack on the computational conception of mind. *Philosophy and Phenomenological Research*, **59**, 41–69.

BRINGSJORD, S. (forthcoming). Creativity, the turing test, and the (better) lovelace test. *Minds and Machines.*

BRINGSJORD, S. & FERRUCCI, D. (2000). *Artificial intelligence and literary creativity: inside the mind of brutus, a storytelling machine.* Mahwah, NJ: Erlbaun.

BRINGSJORD, S. & 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. & ZENZEN, M. (1997). Congnition is not computation: the argument from irreversibility? *Synthese*, *113*, 285–320.

BRINGSJORD, S. & ZENZEN, M. (2001). *Super minds: a defense of uncomputable cognition.* Dordrecht: Kluwer Academic.

DARWIN, C. (1859). *On the origin of the species by means of natural selection.* London: Murray.

DENNETT, D.C. (1995). *Darwin's dangerous idea.* New York: Simon & Sluster.

DESMOND, A. & MOORE, J. (1991). *Darwin.* London: Michael Joseph.

DICKMANN, M.A. (1975). *Large infinitary-languages.* Amsterdam: North-Holland.

EBBINGHANS, H.D., FLUM, J. & THOMAS, W. (1984). *Mathematical logic.* New York: Springer-Verlag.

LEWIS, H. & PAPADIMITRIOU, C. (1981). *Elements of the theory of computation.* Englewood Cliffs, NJ: Prentice Hall.

PELLETIER, F.J. (1986). Seventy five problems for testing automatic theorem provers. *Bell System Technical Journal, 2*, 191–216.

PENROSE, R. (1989). *The emperor's new mind.* Oxford: Oxford University Press.

PENROSE, R. (1994). *Shadows of the mind.* Oxford: Oxford University Press.

PINKER, S. (1994). *The language instinct.* New York: Morront.

PINKER, S. (1997). *How the mind works.* New York: Norton.

PLANTINGA, A. (1993). *Warrant and proper function.* Oxford: Oxford University Press.

RICHARDS, R.J. (1987). *Darwin and the emergence of evolutionary theories of mind and behavior.* Chicago: University of Chicago Press.

RUSSELL, S. & NORVIG, P. (1994). *Artificial intelligence: a modern approach.* Saddle River, NJ: Prentice Hall.

SEARLE, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, **3**, 417–424.

SMULLYAN, R. (1992). *Gödel's incompleteness theorems.* Oxford: Oxford University Press.

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.* London: Keagan Paul.