Ethical Robots: The Future Can Heed Us*

Selmer Bringsjord:
http://www.rpi.edu/~brings
Department of Cognitive Science
Department of Computer Science
Rensselaer AI & Reasoning (RAIR) Lab:
http://www.cogsci.rpi.edu/research/rair/index.php
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu

Abstract

Bill Joy's deep pessimism is now famous. "Why The Future Doesn't Need Us," his defense of that pessimism, has been read by, it seems, *everyone* — and many of these readers, apparently, have been converted to the dark side, or rather more accurately, to the future-is-dark side. Fortunately (for us; *un*fortunately for Joy), the defense, at least the part of it that pertains to AI and robotics, fails. Ours may be a dark future, but we can't know that on the basis of Joy's reasoning. On the other hand, we ought to fear a good deal more than fear itself: we ought to fear not robots, but what some of us may *do* with robots.

Introduction

Bill Joy's deep pessimism is now famous. "Why The Future Doesn't Need Us,"[1] his defense of that pessimism, has been read by, it seems, *everyone* — and a goodly number of these readers, apparently, have been converted to the dark side, or rather, more accurately, to the future-is-dark side. Fortunately (for us; *un*fortunately for Joy), his defense, at least the part of it that pertains to AI and robotics, fails. The arguments he gives to support the view that an eternal night is soon to descend upon the human race because of future robots are positively anemic. Ours may be a dark future, but we can't know that on the basis of Joy's reasoning.

Joy fears a trio: G - N - R, as he abbreviates them: **g**enetics, **n**anotechnology, and **r**obots. I confess to knowing not a bit about G, and I know just enough about N to get myself in trouble by speculating in public about it. I therefore restrict my attention to R: I'm concerned, then, with whether it's rational to believe that Joy's black night will come in large part because of developments in and associated with robotics. For ease of reference, let's lay the relevant proposition directly on the table; I'm concerned with whether the following proposition is established by Joy's reasoning.

$\bar{\mathcal{H}}$ In the relatively near future, and certainly sooner or later, the human species will be destroyed by advances in robotics technology that we can foresee from our current vantage point, at the start of the new millennium.

Let's turn now to the three arguments Joy gives for this proposition, and refute each. Once that's accomplished, we'll end by briefly taking note of the fact that while Joy's techno-fatalism is unfounded, we ought nonetheless to fear a good deal more than fear itself: we ought to fear not robots, but what some of us may *do* with robots.

Argument #1: The Slippery Slope

For his first argument, Joy affirms part of the Unabomber's Manifesto (which appeared in *The Washington Post*, and led to his capture). The argument is quoted and affirmed not only by Joy, but also by Raymond Kurzweil (in his *The Age of Spiritual Machines* (Kurzweil 2000)). Here's the argument:

We — to use the Unabomber's words — "postulate that the computer scientists succeed in developing intelligent machines that can do all things better than human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary." From here, we are to infer that there are two alternatives: the machines are allowed to make their decisions autonomously, without human oversight; or human control is retained. If the former possibility obtains, humans will lose all control, for before long, turning the machines off will end the human race (because by that time, as the story goes, our very metabolisms will be entirely dependent upon the machines). On the other hand, if the latter alternative materializes, "the machines will be in the hands of a tiny elite — just as it is today, but with two differences. Due to improved techniques the elite will have greater control over the masses; and because human work will no longer be necessary the masses will be superfluous, a useless burden on the system." In this scenario, the Unabomber tells us, the elite may decide either to exterminate the masses, or to essentially turn them into the equivalent of domestic animals. The conclusion: if AI continues, humans are doomed. We ought therefore to halt the advance of AI.

*Thanks are due to Konstantine Arkoudas, Paul Bello, and Yingrui Yang for discussions related to the issues treated herein. Special thanks are due to Bettina Schimanski for her robotics work on PERI, and for helping to concretize my widening investigation of robot free will by tinkering with real robots.

[1]The paper originally appeared in *Wired* as (Joy 2000), and is available online: http://www.wired.com/wired/archive/8.04/joy.html. I quote in this paper from the online version, and therefore don't use page numbers. The quotes are of course instantly findable with search over the online version.

Joy quotes the argument in its idiotic entirety. (For the full text, see the Appendix, the final section of the present paper.) He apparently believes that the conclusion is (essentially) $\bar{\mathcal{H}}$, and that the argument is sound. Now, a sound argument is both *formally valid* (the inferences conform to normatively correct rules of inference; i.e., the argument is certified by formal logic) and *veracious* (its premises are true). Unfortunately, not only was the Unabomber a criminal, and insane; he was also a very bad reasoner — and ditto, with all due respect, for anyone who finds his reasoning compelling. This is so because his argument is not only demonstrably invalid, but it also has premises that are at best controversial. (Perhaps at one point during his mathematics career, the Unabomber's brain was working better, but personally, I have my doubts.) Proof by cases (or disjunctive syllogism, or — as it's called in the "proof" given below in Figure 1 — disjunction elimination, or just $\vee$ Elim) is an ironclad rule of inference, of course. If we know that some disjunction

$$P_1 \vee P_2 \vee \cdots \vee P_n$$

holds, and (say) that each P_i leads to proposition Q, then we can correctly infer Q. Because the Unabomber's argument follows the $\vee$ Elim structure, it has an air of plausibility. The structure in question looks like this (where our $\bar{\mathcal{H}}$ is represented here by just H, M stands for the "postulate" in question (the conjunction that intelligent machines will exceed us in all regards, and no human effort will be expended for anything), A for the scenario where the machines make their decisions **a**utonomously, and C for the state of affairs in which humans retain **c**ontrol:

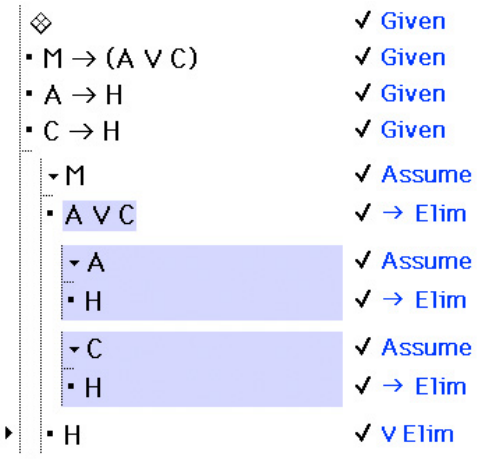

Figure 1: Unabomber Argument Analyzed in Natural Deduction Format

If you look carefully, you'll see that the conclusion of this argument isn't the desired-by-Joy $\bar{\mathcal{H}}$. The conclusion is rather that $\bar{\mathcal{H}}$ follows from M, i.e., M $\rightarrow \bar{\mathcal{H}}$. In order to derive $\bar{\mathcal{H}}$ it of course isn't enough to *suppose* M; instead, M has to be a given; it has to be true, pure and simple. The Unabomber's argument is thus really fundamentally this structure:

> If science policy allows science and engineering in area X to continue, then it's possible that state of affairs P will result; if P results, then disastrous state of affairs Q will possibly ensue; therefore we ought not to allow X.

You don't have to know any formal logic to realize that this is an insanely fallacious pattern. In fact, if this pattern is accepted, with a modicum of imagination you could prohibit any science and engineering effort whatsoever. You would simply begin by enlisting the help of a creative writer to dream up an imaginative but dangerous state of affairs P that is possible given X. You then have the writer continue the story so that disastrous consequences of P arrive in the narrative, and lo and behold you have "established" that X must be banned.

Now of course some of my listeners will have no complaints about M; they will cheerfully affirm this proposition. Given that Turing in 1950 predicted with great confidence that by the year 2000 his test would be passed by our computing machines, while the truth of the matter is that five years into the new millennium a moderately sharp toddler can outthink the smartest of our machines, you'll have to forgive me if I resist taking M as a postulate. To put something in that category, I'm a good deal more comfortable with the kinds of postulates Euclid long ago advanced. Now *they* are plausible.

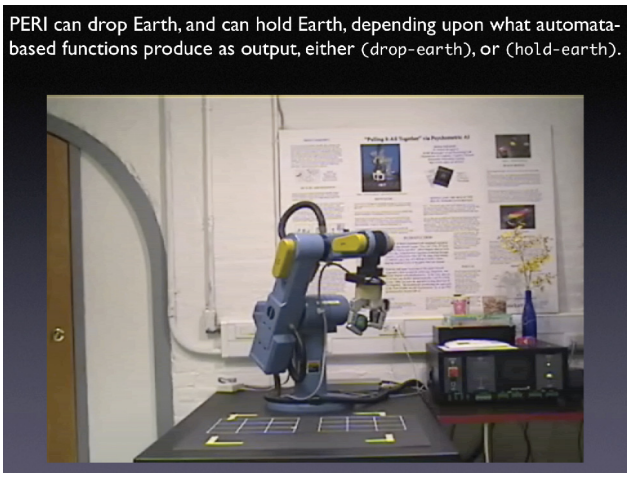

Figure 2: PERI Under the Control of a Finite State Transition Network

Part of my specific discomfort with M is that it's supposed to entail that robots have autonomy. I very much doubt that robots can have this property, in anything like the sense corresponding to the fact that, at the moment, I can decide whether to keep typing, or head downtown and grab a bite to eat, and return to RPI thereafter. Of course, I haven't the space to defend my skepticism. I'll point out only that not many AI researchers have written about this issue, but that John McCarthy has (McCarthy 2000). The odd thing is that his proposal for free will in robots seems to *exclude* free will, in any sense of the term we care about in the human sphere. In his first possibility, free will in robots is identified with *can* in the sense that if a network of intertwined finite state automata were changed, different actions on the part of the

sub-automata would be possible; so it "can" perform these actions. Working with Bettina Schimanski, I have considered the concrete case of PERI, a robot in our lab, dropping or not dropping a ball (which is a miniature earth: dropping is thus "immoral") based on whether the Lisp code that implements the finite state automata in question instructs him to drop or not drop (see Figure 2).[2] It would seem that, in this experiment, whether PERI drops or doesn't is clearly up to *us*, not him. In a second experiment, we took up McCarthy's second suggestion for robotic free will, in which actions performed correspond to those that are provably advisable, where 'provable' is fleshed out with help from standard deduction over knowledge represented in the situation calculus. Here again, I'm mystified as to why anyone would say that PERI is free when his actions are those proved to be advisable. It's not up to him what he does: he does what the prover says to do, and humans built the prover, and set up the rules in question. Where's the autonomy? In general, I can't see how, from a concrete engineering perspective, autonomous robots can be built. Someone might say that randomness is essential, but if whether PERI holds or drops the ball is determined by a random event (see Figure 3), then obviously it's not up to *him* whether the ball is dropped or not.

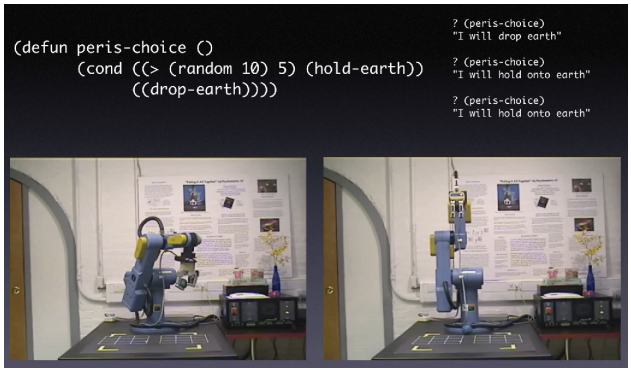

Figure 3: PERI At the Mercy of (Pseudo)Randomness (via Common Lisp's `random`)

Argument #2: Self-Replicating Robots

Joy's second argument is amorphous. He writes:

> Accustomed to living with almost routine scientific breakthroughs, we have yet to come to terms with the fact that the most compelling 21st-century technologies — robotics, genetic engineering, and nanotechnology — pose a different threat than the technologies that have come before. Specifically, robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate.

Unfortunately, though it's clear he's afraid of self-replicating robots, Joy doesn't ever tell us *why* he's afraid. We know, of course, that self-replicating machines (at the level of Turing machines) are quite possible; we've known this since at least Von Neumann (Neumann 1966). Why is it that 40 plus years later that which Von Neumann discovered is so worrisome? What's the new threat? Is it that some company in the business of building humanoid robots is going to lose control of its manufacturing facility, and the robots are going to multiply out of control, so that they end up squeezing us out of our office buildings, crushing our houses and our cars, so that we race to higher ground as if running from a flood? It sounds like a B-grade horror movie. I really do hope that Joy has something just a tad more serious in mind. But what?

I don't know. I could speculate, of course. Perhaps, for example, Joy is worried about the self-replication of very small robots, nano-sized ones. This crosses over from category R to category N, and as you'll recall I said at the outset that I'd refrain from commentary on the supposed dangers of N. I will say only that if something in this direction is what Joy is afraid of, the fact still remains that he doesn't tell us *why* he's afraid. We're just left wondering.

Argument #3: Speed + Thirst for Immortality = Death

This argument goes approximately like this: Humans will find it irresistible to download themselves into robotic bodies, because doing so will ensure immortality (or at least life as long as early Old Testament days). When this happens (and Moore's Law, that magically efficacious mechanism, will soon enough see to it that such downloading is available), the human race will cease to exist. A *new* race, a race of smart and durable machines, will supersede us. And indeed the process will continue *ad indefinitum*, because when race R_1, the one that directly supplants ours, realizes that they can extend their lives by downloading to even more long-lived hardware, they will take the plunge, and so to R_2, and $R_3, \ldots$ we go. Joy writes:

> But because of the recent rapid and radical progress in molecular electronics — where individual atoms and molecules replace lithographically drawn transistors — and related nanoscale technologies, we should be able to meet or exceed the Moore's law rate of progress for another 30 years. By 2030, we are likely to be able to build machines, in quantity, a million times as powerful as the personal computers of today — sufficient to implement the dreams of Kurzweil and Moravec.

Please note that the dreams here referred to are precisely those of achieving virtual immortality on the shoulders of robotic hardware, after shedding the chains of our frail bodies. I'm sure many of my readers will have read Moravec's

[2]The presentation can be found without videos at http://kryten.mm.rpi.edu/PRES/CAPOSU0805/sb_robotsfreedom.pdf. Those able to view keynote, which has the videos of PERI in action embedded, can go to http://kryten.mm.rpi.edu/PRES/CAPOSU0805/sb_robotsfreedom.key.tar.gz. A full account of PERI and his exploits, which haven't until recently had anything to do with autonomy (PERI has been built to match human intelligence in various domains; see e.g. (Bringsjord & Schimanski 2003; Bringsjord & Schimanski 2004)), can be found at http://www.cogsci.rpi.edu/research/rair/pai.

description of his dream, shared in (Moravec 1999). Here's how the argument looks, put more explicitly:

Argument 3, Explicit

(1) Advances in robotics, combined with Moore's Law, will make it possible in about 30 years for humans to download themselves out of their bodies into more durable robotic brains/bodies.

(2) Humans will find this downloading to be irresistible.

$\therefore$ (3) $\bar{\mathcal{H}}$ = In about 30 years, humans will cease to exist as a species.

What are we to say about this argument? Well, it's no more impressive than its predecessors; if a student in an introductory philosophy class, let alone an introductory logic class, submitted this argument, he or she would be summarily flunked. As to formal validity, it fails — but it's no doubt enthymematic. One of the hidden premises is that

(4) If this downloading takes place, humans will cease to exist as a species.

which seems plausible enough. At any rate, I concede that the reasoning could be tidied up to reach formal validity. The real problem is veracity. Why should we think that (1) and (2) hold?

If premise (1) is true, then the human mind must consist wholly in computation; we briefly visited this idea above. Now let's spend a bit more time considering the idea. First, let's get the label straight: if (1) is true, then the doctrine often called *computationalism* is true.

Propelled by the writings of innumerable thinkers (Peters 1962; Barr 1983; Fetzer 1994; Simon 1980; Simon 1981; Newell 1980; Haugeland 1985; Hofstadter 1985; Johnson-Laird 1988; Dietrich 1990; Bringsjord 1992; Searle 1980; Harnad 1991), computationalism has reached every corner of, and indeed energizes the bulk of, contemporary AI and cognitive science. The view has also touched nearly every major college and university in the world; even the popular media have, on a global scale, preached the computational conception of mind. Despite all this, despite the fact that computationalism has achieved the status of a Kuhnian paradigm, the fact is that the doctrine is maddeningly vague. Myriad one-sentence versions of this doctrine float about; e.g.,

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

We don't have the time, today, to sort all this out. Presumably we all have at least some workable grasp of what the doctrine amounts to.[3] The problem for Joy far exceeds the vagueness of the doctrine. The problem is that a refutation of the doctrine has been the conclusion of many deductive arguments. Many of these arguments are ones I've given. (The most recent one recently appeared in *Theoretical Computer Science* (Bringsjord & Arkoudas 2004).) This isn't the place to rehearse these arguments. The point, for now, is simply that they exist, and in light of that, Joy can't just *assume* computationalism.

Now it might be said on Joy's behalf that he doesn't just baldly assume computationalism; instead (so the story goes) he derives this doctrine from Moore's Law, and the fact that tomorrow's computing power will dwarf today's. Unfortunately, here Joy is once more crippled by fallacious reasoning. This is easy to see: Let f be a function from the natural numbers to natural numbers. Now suppose that the storage capacity and speed of today's computers grows for 1,000 years at rates that exceed even what Joy has in mind; and suppose, specifically, that C is the best computer available in 3005. Question: Does it follow that C can compute f? No, of course not, and the proof is trivial: Simply define $f(n)$ to be the maximum productivity of n-state Turing machines with alphabet $\{0, 1\}$, where these machines are invariably started on an empty tape, and their productivity corresponds to the number of contiguous 1s they leave on the tape, after halting with their read/write head on the leftmost of these 1s. Since this famous function, the so-called Σ or "busy beaver" function, is Turing-uncomputable (Boolos & Jeffrey 1989), C, no matter how fast, can't compute f. (Of course, any Turing-uncomputable function will do. E.g., the halting problem would do just fine.) The problem is that Joy suffers from some sort of speed fetish; I've written about this fetish elsewhere (Bringsjord 2000). Speed is great, but however fast standard computation may be, it's still by definition at or below the Turing Limit. It doesn't follow from Moore's Law that human mentation can be identified with the computing of functions at or below this limit. There are a lot more functions above this limit than below it, and it may well be that some of the functions we process are in this space. In fact, I've written a book in defense of just this possibility (Bringsjord & Zenzen 2003).

The amazing thing to me is that we in AI *know* that speed isn't a panacea. Does anyone seriously maintain that the bottleneck in natural language processing is due to the fact that computers aren't fast enough? No matter how fast the hardware you're programming may be, to program it to compute g you need to know what g is. We don't seem to know what the function is that underlies, say, our ability to learn language, to use it to give a lecture, and to debate, afterwards, those who heard it and didn't buy it. (I know none of you

[3] This is as good a place as any to point out that, as the parentheticals associated with a number of the propositions on the list just given indicate, by the lights of some computationalists we aren't pure software, but are embodied creatures. Joy and Moravec (and Hillis) assume that human persons are in the end software that can be attached to this or that body. That seems like a pretty big assumption.

here are in that category.)

Argument 3, Explicit has another vulnerable premise: (2). Is it really true that humans would take up offers to be re-embodied as robots? Suppose I came to you and said: "Look, you're going to die soon, because your body is going to give out. It might not happen tomorrow, but it will next week, or next month, or in a few years. Now, see this robot over here?" I point to a glistening humanoid robot. "I'll tell you what I'll do. You sit in this chair over here. It'll scan your brain and decipher the code that makes you you. Once this code is extracted, we'll vaporize your old-style body, and download you into the robot here. You'll live a lot longer, hundreds of years longer. And as an added bonus, I'll agree contractually that when this robot starts to fail, my descendants will jump you to an even more durable robotic body."

I'm not sure I find this offer irresistible.[4] How about you?

More to Fear than Fear

Unfortunately, Joy unwittingly alludes to something we *should* fear. It's not robotics; nor is it the other pair in GNR. We need to fear *us* — or at least some of us. We need to fear those among us with just enough brain power to use either G or N or R as a weapon. As Joy writes:

> Thus we have the possibility not just of weapons of mass destruction but of knowledge-enabled mass destruction (KMD), this destructiveness hugely amplified by the power of self-replication. I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals.

Mosquitoes replicate, as do a thousand thousand other pests. *Ceteris paribus*, robots, whether big or small, at least as I see it, will be at worst pests when left to their own devices. But some humans will no doubt seek to use robots (and for that matter softbots) as weapons against innocent humans. This is undeniable; we can indeed sometimes see the future, and it does look, at least in part, very dark. But it won't be the robots who are to blame. *We* will be to blame. The sooner we stop worrying about inane arguments like those Joy offers, and start to engineer protection against those who would wield robots as future swords, the better off we'll be.

Appendix

The full quote of the Unabomber's fallacious argument, which appears also in Joy's piece:

First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better than human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. Either of two cases might occur. The machines might be permitted to make all of their own decisions without human oversight, or else human control over the machines might be retained.

If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all the power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide.

On the other hand it is possible that human control over the machines may be retained. In that case the average man may have control over certain private machines of his own, such as his car or his personal computer, but control over large systems of machines will be in the hands of a tiny elite — just as it is today, but with two differences. Due to improved techniques the elite will have greater control over the masses; and because human work will no longer be necessary the masses will be superfluous, a useless burden on the system. If the elite is ruthless they may simply decide to exterminate the mass of humanity. If they are humane they may use propaganda or other psychological or biological techniques to reduce the birth rate until the mass of humanity becomes extinct, leaving the world to the elite. Or, if the elite consists of soft-hearted liberals, they may decide to play the role of good shepherds to the rest of the human race. They will see to it that everyone's physical needs are satisfied, that all children are raised under psychologically hygienic conditions, that everyone has a wholesome hobby to keep him busy, and that anyone who may become dissatisfied undergoes "treatment" to cure his "problem." Of course, life will be so

[4] Any kind of reassurance would require that that which it feels like to be me had been reduced to some kind of third-person specification — which many have said is impossible. I've alluded above to the fact that today's smartest machines can't verbally out-duel a sharp toddler. But at least we do have computers that can understand *some* language, and we continue to press on. But we are really and truly nowhere in an attempt to understand consciousness in machine terms.

purposeless that people will have to be biologically or psychologically engineered either to remove their need for the power process or make them "sublimate" their drive for power into some harmless hobby. These engineered human beings may be happy in such a society, but they will most certainly not be free. They will have been reduced to the status of domestic animals.

References

[Barr 1983] Barr, A. 1983. Artificial intelligence: Cognition as computation. In Machlup, F., ed., *The Study of Information: Interdisciplinary Messages*. New York, NY: Wiley-Interscience. 237–262.

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

[Bringsjord & Arkoudas 2004] Bringsjord, S., and Arkoudas, K. 2004. The modal argument for hypercomputing minds. *Theoretical Computer Science* 317:167–190.

[Bringsjord & Schimanski 2003] Bringsjord, S., and Schimanski, B. 2003. What is artificial intelligence? psychometric AI as an answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI–03)*, 887–893.

[Bringsjord & Schimanski 2004] Bringsjord, S., and Schimanski, B. 2004. 'pulling it all together" via psychometric ai. In *Proceedings of the 2004 Fall Symposium: Achieving Hman -Level Intelligence through Integrated Systems and Research*, 9–16.

[Bringsjord & Zenzen 2003] Bringsjord, S., and Zenzen, M. 2003. *Superminds: People Harness Hypercomputation, and More*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

[Bringsjord 2000] Bringsjord, S. 2000. A contrarian future for minds and machines. *Chronicle of Higher Education* B5. Reprinted in *The Education Digest* **66.6**: 31–33.

[Dietrich 1990] Dietrich, E. 1990. Computationalism. *Social Epistemology* 4(2):135–154.

[Fetzer 1994] Fetzer, J. 1994. Mental algorithms: Are minds computational systems? *Pragmatics and Cognition* 2.1:1–29.

[Harnad 1991] Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1(1):43–54.

[Haugeland 1985] Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

[Hofstadter 1985] Hofstadter, D. 1985. Waking up from the Boolean dream. In *Metamagical Themas: Questing for the Essence of Mind and Pattern*. New York, NY: Bantam. 631–665.

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

[Joy 2000] Joy, W. 2000. Why the Future Doesn't Need Us. *Wired* 8(4).

[Kurzweil 2000] Kurzweil, R. 2000. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York, NY: Penguin USA.

[McCarthy 2000] McCarthy, J. 2000. Free will–even for robots. *Journal of Experimental and Theoretical Artificial Intelligence* 12(3):341–352.

[Moravec 1999] Moravec, H. 1999. *Robot: Mere Machine to Transcendant Mind*. Oxford, UK: Oxford University Press.

[Neumann 1966] Neumann, J. V. 1966. *Theory of Self-Reproducing Automata*. Illinois University Press.

[Newell 1980] Newell, A. 1980. Physical symbol systems. *Cognitive Science* 4:135–183.

[Peters 1962] Peters, R. S., ed. 1962. *Body, Man, and Citizen: Selections from Hobbes' Writing*. New York, NY: Collier.

[Searle 1980] Searle, J. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417–424.

[Simon 1980] Simon, H. 1980. Cognitive science: The newest science of the artificial. *Cognitive Science* 4:33–56.

[Simon 1981] Simon, H. 1981. Study of human intelligence by creating artificial intelligence. *American Scientist* 69(3):300–309.