VOLUME 22 NUMBER 4
April 2008

AI & SOCIETY

The Journal of Human-Centred Systems

Springer

AI & SOCIETY

Editorial Board

Advisory Board

AI & SOCIETY

Volume 22 . Number 4 . April 2008

Special Issue: Ethics and artificial agents
Guest Editor: Steve Torrance

Springer

AI & SOCIETY

Aims and scope: *AI & Society* is an International JDumal, publishing refereed scholarly articles, position papers, debates, short communications and reviews. Established in 1987, the journal focuses on the issues of policy, design, applications of information, communications and new media technologies, with a particular emphasis on cultural, social, cognitive, economic, ethical and philosophical implications.

AI & Society is broad based and strongly interdisciplinary, It provides an international forum for "over the horizon" analysis of the gaps and possibilities of rapidly evolving 'knowledge society', with a hmnanistic vision of society, culture and technology.

Subscription information
ISSN print edition 0951-5666
ISSN electronic edition 1435-5655

• The Americas (North, South, Central America and the Caribbean)
joumals-ny@springer.com

• Outside the Americas
subscriptions@springer.com

• United Kingdom
Please ask for the appropriate subscription rate in £. Prices are net prices subject to local VAT plus carriage charges. Orders and claims shall be directly sent to: Springer-Verlag London Ltd. Ashbourne House, The Guildway, Old Portsmouth Road, Artington, Guildford, GU3 ILP, UK. FREECALL from UK Tel. *00800* SPRINGER (77746437).

• Orders and inquiries
The Americas (North, South, Central America and the Caribbean)
Springer, Journal Fulfillment
P.O. Box 2485, Secaucus, NJ *07096,* USA
Tel.: + 1-800-SPRINGER, + 1-201-348-4033
or + 1-212-4601500, Fax: + 1-201,3484505
e-mail: journals-ny@springer.com

• Outside the Americas
via a bookseJler or
Springer Distribution Center GmbH
Haberstrasse 7, 69126 Heidelberg, Gennany
Tel.: +49-6221-345-4303, Fax: *+49-6221-3454304*
e-mail: subscriptions@springer.com
Business hours: Monday to Friday
8 a,m, to 8 p.m, local time and on Gennan public holidays.

Cancellations must be received by September 30 to take effect at the end of the same year.

Changes of address. Allow six weeks for all changes to become effective, All communications should include both old and new addresses (with postal codes) and should be accompanied by a mailing label from a recent issue,

According to 4 Sect. 3 of the Gennan Postal Services Data Protection Regulations, if a subscriber's address changes, the German Post Office can inform the publisher of the new address even if the subscriber has not submitted a formal application for mail to be forwarded. Subscribers not in agreement with this procedure may send a written complaint to Customer Service Joumals, within 14 days of publication of this issue.

Back volumes. Prices are available on request.

Microform editions are available from: ProQuest. Further information available at: http://www.iLproquest.com/umi/

Electronic edition: An electronic edition of this journal is available at springerlink,com .

Science communication
Beverley Ford
Executive Editor, Computer Science Springer, Ashbourne House, The Guildway, Old Portsmouth Road,
Guildford, Surrey, GU3 ILP, UK
Tel.: +44 1483 734646
Fax: + 44 1483 734411
E-mail: beverley.ford@springer.com

Production
Springer, Helmut Petri, Journal Production Computer Science
Postfach 105280, *69042* Heidelberg, Germany
E-mail: Helmut.Petri@springer.com
TeL/Fax: + 49-6221-487-8494/68494

Address for courier, express and registered mail:
Tiergartenstrasse 17, 69121 Heidelberg, Germany

Typesetters: Scientific Publishing Services Pvt. Ltd., Chennai, India

Printers: Krips, Meppel, The Netherlands
Printed on acid-free paper
Springer London Limited
is a part of Springer Science + Business Media

springer.com

AI & Soc (2008) 22:539–550
DOI 10.1007/s00146-007-0090-9

ORIGINAL ARTICLE

Ethical robots: the future can heed us

Selmer Bringsjord

Received: 10 January 2006 / Accepted: 2 February 2007 / Published online: 13 March 2007

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; unfortunately 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 cannot 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; unfortunately 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

[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.

S. Bringsjord (✉)
Department of Cognitive Science, Department of Computer Science, Rensselaer Polytechnic Institute (RPI), Rensselaer AI and Reasoning (RAIR) Lab, Troy, NY 12180, USA
e-mail: selmer@rpi.edu
URL: http://www.rpi.edu/~brings; http://www.cogsci.rpi.edu/research/rair

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 cannot know that on the basis of Joy's reasoning.

Joy fears a trio: G–N–R, as he abbreviates them: genetics, nanotechnology, and robots. 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 am concerned, then, with whether it is 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 us lay the relevant proposition directly on the table; I am concerned with whether the following proposition is established by Joy's reasoning.

{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 us turn now to the three arguments Joy gives for this proposition, and refute each. Once that is accomplished, we will 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 No. 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 is 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.

Joy quotes the argument in its entirety. (For the full text, see the Appendix, the final section of the present paper.) He apparently believes that the conclusion is (essentially) {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 is called in the "proof" given below in Fig. 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 {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 autonomously; and C for the state of affairs in which humans retain control:

If you look carefully, you will see that the conclusion of this argument is not H (i.e., the desired-by-Joy {H}). The conclusion is rather that H follows from M, i.e., M $\rightarrow$ H. In order to derive H it of course is not 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 is possible that state of affairs M will result; if M results, then either

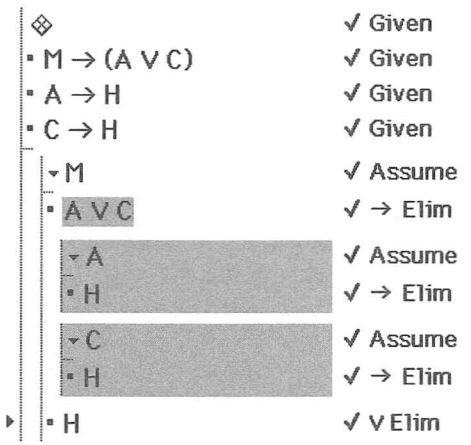

Fig. 1 Unabomber argument analyzed in natural deduction format

disastrous state of affairs *A* or disastrous state of affairs C will ensue. We ought not allow work in science and engineering areas to continue if doing so has disastrous results; therefore we ought not to allow work in area *X* to continue.

You do not have to know any formal logic to realize that this is a 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 10 and behold you have "established" that *X* must be banned.

Now of course some of my readers 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 (see Turing 1950), 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 will have to forgive me if I resist taking M as a postulate. To put something in that category, I am a good deal more comfortable with the kinds of postulates Euclid long ago advanced. Now *they* are plausible.

Part of my specific discomfort with M is that it is 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 have not the space to defend my skepticism. I will 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-automaton 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 automaton in question instructs him to drop or not drop (see Fig. 2).[2] It would seem that, in this experiment, whether PERI drops or does not 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

2 The presentation can be found without videos at http://www.kryten.mm.rpi.edu/PRES/CAPOSD0805/sb_robotsfreedom.pdf. Those able to view keynote, which has the videos of PERI in action embedded, can go to http://www.kryten.mm.rpi.eduIPRES/CAPOSD0805/sb_robotsfreedom.key.tar.gz. A full account of PERI and his exploits, which have not until recently had anything to do with autonomy (PERI has been built to match human intelligence in various domains; see e.g., Bringsjord and Schimanski 2003, 2004) can be found at http://www.cogsci.rpi.edu/research/rair/pai.

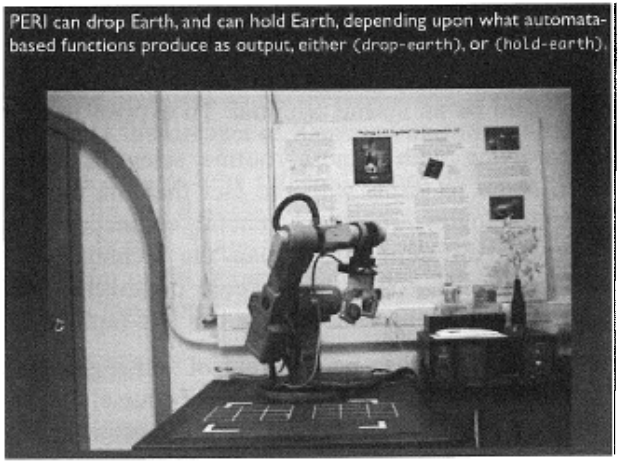

Fig. 2 PERI under the control of a finite state transition network

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 am mystified as to why anyone would say that PERI is free when his actions are those proved to be advisable. It is 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 is the autonomy? In general, I cannot 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 Fig. 3), then obviously it is not up to *him* whether the ball is dropped or not. At any rate, the onus is clearly on anyone claiming that robots can have human-like autonomy, given that no such robot has been built, or even designed. Finally, from an engineering perspective, we have good reason to believe that the nature of robots, as artifacts programmed by us, suggests that they are human-controllable (see, e.g., Arkoudas and Bringsjord 2005).

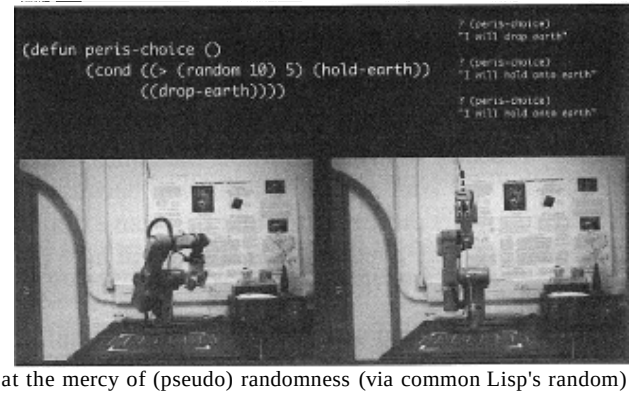

Fig. 3 PERI at the mercy of (pseudo) randomness (via common Lisp's random)

Argument No.2: self-replicating robots

Joy's second argument is amorphous, but at least has the virtue of not resting on reasoning produced by an insane criminal. To express it, 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 nano-technology-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 is clear he is afraid of self-replicating robots, Joy does not ever tell us *why* he is afraid. We know, of course, that self-replicating machines (at the level of Turing machines) are quite possible; we have known this since at least Von Neumann (1966). Why is it that 40 plus years later, that which Von Neumann discovered is so worrisome? What is 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, and, terror-stricken, 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 do not 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 will recall I said at the outset that I had to 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 does not tell us *why* he is afraid. We are just left wondering.

Argument No.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, ... 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 am sure many of my readers will have read Moravec's description of his dream, shared in (Moravec 1999). Here is how the argument looks, put more explicitly:

Argument No.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.
3. *H* = In about 30 years, humans will cease to exist as a species.

What are we to say about this argument? Well, it is 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 on the other hand it is 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 us spend a bit more time considering the idea. First, let us 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, 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.

I do not have the space here to sort all this out. Presumably most readers 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 have given.[4] This is not the place to rehearse these arguments.[5] The point, for now, is simply that they exist, and in light of that, Joy cannot just assume computationalism.

Now it might be said on Joy's behalf that he does not 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 grow 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 3006. 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 is they leave on the tape, after halting with their read/write head

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 are not 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.

4 The most recent one appeared in *Theoretical Computer Science* (Bringsjord and Arkoudas 2004). For a formal list that was up-to-date as of 2003, and reached back to my *What Robots Can and Cannot Be* (1992), see my *Superminds (2003).*

5 However, it does occur to me that it would perhaps be nice if a new argument against computationalism could be introduced in the present paper. Accordingly, here is one such argument, one that happens to be in line with the themes we are reflecting upon herein: Argument NO.4. 1. If computationalism is true, then concerted, global efforts undertaken by the world's best relevant scientists and engineers to build computing machines with the cognitive power of human beings will succeed after *n* years of effort-the "clock" having been started in 1950. 2. Concerted, global efforts undertaken by the world's best relevant scientists and engineers to build computing machines with the cognitive power of human beings have not succeeded after *n* years of effort (the clock). 3. Computationalism is false. Obviously, my case against Joy hinges not a bit on this argument. But I do think this argument should give pause to today's computationalists. I have made it a point to ask a number of such strong "believers" how many years of failure would suffice to throw the truth of computationalism into doubt in their minds-and have never received back a number. But clearly, there must exist some *n* for which Argument No.4 becomes sound. It seems to me that 50 is large enough, especially given that we have not produced a machine able to converse at the level of a sharp toddler.

on the leftmost of these Is. Since this famous function, so-called Σ or "busy beaver" function, is Turing-uncomputable (Boolos and Jeffrey 1989), C, no matter how fast, cannot 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 have written about this fetish elsewhere (Bringsjord 2000). Speed is great, but however fast standard computation may be, it is still by definition at or below the Turing Limit. It does not follow from Moore's Law that human mentation can be identified with the computing of functions at or below this limit. There are 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 have written a book in defense of just this possibility (Bringsjord and Zenzen 2003).

The amazing thing to me is that we in AI know that speed is not a panacea. Does anyone seriously maintain that the bottleneck in natural language processing is due to the fact that computers are not fast enough? No matter how fast the hardware you are programming may be, to program it to compute g you need to know what g is. We do not seem to know what the functions are that underlie, say, our ability to learn language, to use it to give a lecture, and to debate, afterwards, those who heard it and did not completely buy it.[6]

Argument No.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 am not sure I find this offer irresistible.[7] How about yoU?8

6 Let me point out here that it is entirely possible to do some first-rate thinking predicated on the supposition that human-level robots will eventually arrive. At the conference where I presented the keynote lecture centered around an ancestor of the present paper, such thinking was carried out by Torrance (2005) and Moor (2005).

7 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 have alluded above to the fact that today's smartest machines cannot 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.

8 Of course, some philosophers (e.g., Parfit 1986) have championed views of personal identity that seem to entail the real possibility of such downloading. But this is quite beside the point on the table, which is whether you would, in my thought-experiment, take the plunge. It is easy enough to describe thought-experiments in which even conservative folks would take the plunge. For example, if you knew that you were going to die in one hour, because an atom bomb is going to be detonated directly below your feet, you might well, out of desperation, give the downloading a shot. But this is a different thought-experiment.

More to fear than fear

Unfortunately, Joy unwittingly alludes to something we should fear. It is not robotics; nor is it the other pair in G-N-R. 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 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 would not 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 will be.

Acknowledgements Thanks are due to Steve Torrance, Michael Anderson, and Jim Moor, for comments and suggestions offered after the keynote presentation that was based on an ancestor of this paper (at AAAI's 2005 fall symposium on machine ethics). Thanks are also 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. Finally, I am grateful to two anonymous referees for comments and suggestions.

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 cannot 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 would not 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 purpose-less 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

Arkoudas K, Bringsjord S (2005) Toward ethical robots via mechanized deontic logic. In: Technical report-machine ethics: papers from the AAAI fall symposium; FS-05-06, American Association of Artificial Intelligence, Menlo Park, CA, pp 24-29

Barr A (1983) Artificial intelligence: cognition as computation. In: Machlup F (eds) The study of information: interdisciplinary messages, Wiley-Interscience, New York, NY, pp 237-262
Boolos GS, Jeffrey RC (1989) Computability and logic. Cambridge University Press, Cambridge, UK
Bringsjord S (1992) What robots can and can't be. Kluwer, Dordrecht, The Netherlands
Bringsjord S (2000) A contrarian future for minds and machines Chronicle of higher education p B5. Reprinted in The Education Digest 66(6):31-33
Bringsjord S, Arkoudas K (2004) The modal argument for hypercomputing minds. Theor Comput Sci 317:167-190
Bringsjord S, 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), San Francisco, CA, pp 887-893
Bringsjord S, Schimanski B (2004) Pulling it all together via psychometric AI. In: Proceedings of the 2004 fall symposium: achieving human-level intelligence through integrated systems and Research, Menlo Park, CA, pp 9-16
Bringsjord S, Zenzen M (2003) Superminds: people harness hypercomputation, and more. Kluwer Academic, Dordrecht, The Netherlands
Dietrich E (1990) Computationalism. Soc Epistemology 4(2):135-154
Fetzer J (1994) Mental algorithms: are minds computational systems? Pragmatics Cogn 2(1):1-29
HarnadS (1991) Other bodies, other minds: a machine incarnation of an old philosophical problem. Minds Mach 1(1):43-54
Haugeland J (1985) Artificial intelligence: the very idea. MIT Press, Cambridge, MA
HofstadterD (1985) Waking up from the Boolean dream. In: metamagical themas: questing for the essence of mind and pattern, Bantam, New York, NY, pp 631-665
Johnson-Laird P (1988) The computer and the mind. Harvard University Press, Cambridge, MA
JoyW (2000) Why the future doesn't need us. Wired 8(4)
Kurzweil R (2000) The age of spiritual machines: when computers exceed human intelligence. Penguin USA, New York, NY
McCarthy J (2000) Free will-even for robots. J Experimental Theor Artif Intell12(3):341-352
Moor J (2005) The nature and importance of machine ethics. In: technical report-machine ethics: papers from the AAAI fall symposium; FS-05-06, American Association of Artificial Intelligence, Menlo Park, CA
Moravec H (1999) Robot: mere machine to transcendant mind. Oxford University Press, Oxford, UK
Neumann JV (1966) Theory of self-reproducing automata. Illinois University Press, IL
Newell A (1980) Physical symbol systems. Cogn Sci 4:135-183
Parfit D (1986) Reasons and persons. Oxford University Press, Oxford, UK
Peters RS (ed) (1962) Body, man, and citizen: selections from hobbes' writing. Collier, New York, NY
Searle J (1980) Minds, brains, and programs. Behav and Brain Sci 3:417-424
Simon H (1980) Cognitive science: the newest science of the artificial. Cogn Sci 4:33-56
Simon H (1981) Study of human intelligence by creating artificial intelligence. Am Sci 69(3):300-309
Torrance S (2005) A robust view of machine ethics. In: Technical report-machine ethics: papers from the AAAI fall symposium; FS-D5-06, American Association of Artificial Intelligence, Menlo Park, CA
Turing A (1950) Computing machinery and intelligence. Mind LIX 236:433-460