Open Source AI

Bill HIBBARD
University of Wisconsin - Madison

Abstract. Machines significantly more intelligent than humans will require changes in our legal and economic systems in order to preserve something of our human values. An open source design for artificial intelligence (AI) will help this process by discouraging corruption, by enabling many minds to search for errors, and by encouraging political cooperation. The author's experience developing open source software provides anecdotal evidence for the healthy social effects of open source development.

Keywords. Open source, AI ethics, AI politics.

Introduction

There is little doubt that humans will create machines significantly more intelligent than themselves during the twenty first century. Neuroscience is finding a large number of correlations between mental and physical brain behaviors. If brains do not explain minds, then these correlations would be absurd coincidences. And if physical brains do explain minds, then our relentless technology will create minds with artificial physical brains.

Under our current legal and economic systems, super-intelligent machines will create social chaos. They will be able to do every job more efficiently than humans, resulting in 100% unemployment. They will create great wealth for their owners while denying the vast majority any way to support themselves. Technology will enable humans to increase their own intelligence via artificial extensions of their brains. But then each person's intelligence will depend on what sort of brain they can afford. Less intelligent humans will not be able to understand the languages used by the most intelligent humans and machines and thus will be effectively excluded from any meaningful political process.

Most people will object to these social consequences and seek to alleviate them through some adjustment to our legal and economic systems, including constraints on the designs of artificial intelligence (AI). These adjustments ought to be a topic for this workshop.

Of course these issues have been raised. The Hebrew myth of the golem depicts the possible unintended consequences of creating artificial life by magical means. In the early nineteenth century Shelley explored the ethical issues of using technology to create artificial life [1]. Asimov's three laws of robotics were probably the first attempt to define ethical standards for interactions between humans and AI [2, 3]. Vinge applied the term *singularity* to the predicted explosive increase of technology and intelligence when machines more intelligent than humans take over their own development, and described the difficulty of predicting the consequences of this event [4]. Now it is commonplace for science fiction stories and movies to depict intelligent machines as a threat to humans, and the issue of AI ethics has emerged as a serious subject [5, 6, 7]. It is usual to depict the threat as AI versus humanity, but it is also important to counter the threat of AI enabling a small group of humans to take dictatorial control over the rest of humanity.

1. Transparency

Human society could not have achieved the efficiency necessary to create AI without *specialization*, where different people become experts in different types of work and trade the results of their work. In many cases experts act as the *agents* for many other people, representing their interests in important decisions. For example, the laws of society are determined by government experts, hopefully elected by those they represent or at least supervised by elected representatives. Leaders of large corporations act as agents for corporate owners, usually subject to some sort of election. However, whenever one person serves as agent for others, there is the possibility of *corruption*, in which the agent serves his own interests at the expense of those he represents. An essential tool for preventing corruption is *transparency*, in which the decisions and circumstances of agents are made known to those they represent.

The creation of super-human AI is the most difficult challenge in human history and will require extreme expertise. It will also require large resources, controlled by powerful government and corporate leaders. The results of super-human AI will be much more profound for humanity than those of any previous technology. Thus, whether they like it or not, the designers and resource managers for AI represent the interests of all of humanity. Current law and public opinion do not recognize AI designers and managers as humanity's agents, so they may feel no need to represent humanity's interests. Even if they do acknowledge that they represent humanity's interests, the stakes are so high that the temptation for corruption will be intense.

Protecting the interests of humanity will require that law and public opinion change to recognize that AI designers and managers are humanity's agents. Once this agent relation is recognized, preventing corruption will require transparency, which includes an open source design for AI.

2. Many Minds Searching for Errors

Yudkowsky has proposed an effort to produce a design for AI whose friendliness toward humans can be proved as it evolves indefinitely into the future [8]. Legg's blog includes a fascinating debate with Yudkowsky over whether such a proof is possible [9]. Legg produced a proof that it is not possible to prove what an AI will be able to achieve in the physical world, and Yudkowsky replied that he is not trying to prove what an AI can achieve in the physical world but merely trying to prove that the AI maintains friendly intentions as it evolves into the indefinite future. But intentions must be implemented in the physical world, so proving any constraint on intentions requires proving that the AI is able to achieve a constraint on the implementation of those intentions in the physical world. That is, if you cannot prove that the AI will be able to achieve a constraint on the physical world then you cannot prove that it will maintain a constraint on its intentions.

If there can be no guarantee that an AI design remains friendly towards humans as it evolves, the next best approach is for the design to be examined for flaws by as many intelligent humans as possible. An open source design can be studied by anyone who cares to.

Even if a proof of friendliness is possible, proposed mathematical proofs often contain errors that are best found by open publication. And an open source design will enable a large community to verify that the design conforms to the conditions defined in the proof.

3. AI Politics

Politics will be necessary to change law and public opinion to recognize the responsibility of AI designers and managers for general human welfare. The academic debate about whether AI is possible and about the social effects of AI is already strenuous. The general political debate, when it arises, will be even more strenuous. And given the high stakes, it should be. Hopefully the scientific and technical communities will play a major role, with politicians listening to the U.S. National Academies of Science and Engineering and similar institutions internationally.

The transparency of open source design will help set a tone of cooperation, rather than competition, in the political debate. It will reassure the public that they, and experts they trust, are included in the decision process.

The political position of the Singularity Institute for Artificial Intelligence (SIAI) is puzzling. SIAI's long-time leader has expressed contempt for public opinion and politics [10, 11]. But he also contends that AI must be designed according to strict constraints to avoid a disaster for humanity. How are these constraints to be enforced if not by the force of government? Does the SIAI intend to create a friendly AI that takes control of the world before governments have time to react, and before any much better funded competitors can create AI? And given the stated contempt for public opinion and politics, it is hard to understand the considerable efforts of the SIAI to publicize the issues of AI via their Singularity Summits and other venues. In fairness, SIAI leadership is becoming more diffuse. For example, in early 2007 Ben Goertzel became SIAI's Director of Research. Under his leadership SIAI is sponsoring an open source AI project [12]. AI is inevitably becoming a political issue, and it will be interesting to see whether SIAI expresses cooperation rather than hostility toward politics.

4. Cautions

The potential problem with open source AI is that it may enable malicious and incompetent people to build harmful AIs. The answer to this problem is for benign and competent people to out compete the malicious and incompetent in building AI, and for the "good guy" AIs to take up the task of preventing "bad guy" AIs. Liberal democracies are the leaders in science and technology and AI is very likely to first appear in one of these societies. In democracies, the good intentions of government policy generally depend on citizens being informed and engaged on issues. Thus the primary need is for creation of a broad public movement for AI to benefit all people. Given such a movement, government will need the competent advice of leading scientists and engineers, such as provided by the U.S. National Academies.

Some are worried about a hard takeoff scenario, in which the "good guys" don't have time to react between the first human level AI and the explosive increase of intelligence which enables the AI to take control of the world. However, progress in AI has been and will continue to be gradual up to the point of human level AI, giving the public plenty of time to get to know sub-human level AIs in the form of natural language voice user interfaces and robots smart enough to do numerous household and business tasks. These experiences will start many people thinking seriously about what is coming, leading to real political debate. In fact, many politicians are already aware of the coming singularity, but do not discuss it for fear of alienating voters. Even once human level AI is achieved, there may be a time interval of years or even decades before the explosive intelligence increase, because of the time brains must spend learning intelligent behaviors through interactions with the world. Democratic governments control technological resources much greater than any other institution and, given this time interval for political response, will win a race to the singularity based on an open source design.

Nuclear and biological weapons are examples of technologies where humanity has arguably benefited from keeping the details secret. However, those weapons technologies are only useful for killing and harming people. The elites who control their secrets have incentive not to use them, because their use would cause loss of productive capacity and cause the elites to lose legitimacy. In contrast, selfish use of AI by an elite will not cause loss of productive capacity and will enhance the power of the elite (we can assume that an elite with access to a super-intelligent AI will not behave stupidly). Furthermore, AI is a technology that can be applied to prevent others from developing that same technology. This is not nearly so true for nuclear and biological weapons technologies. Thus these weapons technologies are not good analogies for AI.

5. Personal Experiences

My Vis5D, in 1989, was the first open source 3-D visualization system, and my subsequent VisAD and Cave5D visualization systems were also open source. My experience with these systems was that their open source brought out the best in the people who joined their user and developer communities. Whatever fears I had before making Vis5D open source never materialized, whereas I was constantly surprised by the generosity and helpfulness of others. The mailing lists serving these communities had occasional foolishness but never any flame wars. This experience was in sharp contrast with the mistrust and technical problems in the community of a proprietary system in my university department. And I have generally observed this difference of politics between open source and proprietary system communities. While this is all anecdotal, it has been strong enough to reshape my general political attitudes [13].

AI is inevitably becoming a political issue [14]. If the AI development community approaches the public with transparency, including open source designs, the public will surprise many of us with their reasonableness.

References

[1] Shelly, M. 1818. *Frankenstein*. http://www.literature.org/authors/shelley-mary/frankenstein/index.html

[2] Asimov, I. Runaround, Astounding Science Fiction, March 1942.

[3] Asimov, I. 1968. *I, Robot*. London. Grafton Books.

[4] Vinge, V. Vinge, V. 1993. The coming technological singularity. *Whole Earth Review*, Winter issue.

[5] Hibbard, W. Super-Intelligent Machines. Computer Graphics 35(1), 11-13. 2001. http://www.ssec.wisc.edu/~billh/visfiles.html

[6] Bostrom, N. Ethical Issues in Advanced Artificial Intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al.*, Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17. http://www.nickbostrom.com/ethics/ai.html

[7] Goertzel, B. Universal Ethics: The Foundations of Compassion in Pattern Dynamics. October 25, 2004. http://www.goertzel.org/papers/UniversalEthics.htm

[8] Yudkowsky, E. (2006) Knowability of FAI. http://sl4.org/wiki/KnowabilityOfFAI

[9] Legg, S. Unprovability of Friendly AI. September 15, 2006. http://www.vetta.org/?p=6

[10] Yudkowsky, E. CoherentExtrapolatedVolition. http://www.sl4.org/wiki/CollectiveVolition

[11] http://www.ssec.wisc.edu/~billh/g/message10.txt

[12] http://www.opencog.org/

[13] Hibbard, W. Singularity Notes. http://www.ssec.wisc.edu/~billh/g/Singularity_Notes.html

[14] Johnson, G. *Who Do You Trust More: G.I. Joe or A.I. Joe?* New York Times, February 20, 2005.