SYNESIS
A JOURNAL OF SCIENCE, TECHNOLOGY, ETHICS, AND POLICY

Mitigating potential hazards to humans from the development of intelligent machines

William Daley[1]

1. Science and the Public Masters Program, Graduate School of Education, State University of New York at Buffalo, 367 Baldy Hall, Buffalo, NY, USA, Email: wjdaley@buffalo.edu.

Abstract
Artificially intelligent machines and robots grow ever more present and capable. The evolution of intelligent machines may have consequences that are difficult to limit or stop. It is important to examine possible problems now in order to set standards and safeguards to guide the intelligent machine evolution. I suggest that national and international government panels on artificial intelligence issues, similar to the Presidential Commission for the Study of Bioethical Issues, will be necessary to direct the formulation of guidelines that establish such standards for the creation and uses of artificially intelligent machines.

Key words: intelligent machine shutoff, robot hazards, roboethics, artificial intelligence, public policy

The evolution of intelligent machines

As artificially intelligent machines and robots grow ever smarter, their presence in science fiction — sometimes comforting but often menacing — also grows. Although we certainly have no need for concern about machines having anywhere near the mental capabilities of humans at the present time or anytime soon, their evolution continues. While the dark literary and film portrayals are only fiction, the evolution of intelligent machines may have negative consequences that are difficult to limit or stop. Therefore, it is important to address possible problems now in order to set in place standards and safeguards to guide the intelligent machine evolution.

Children and adults alike are fascinated by humanoid machines with which they can interact, and we certainly have many jobs for such machines. NASA has a “Robonaut” aboard the International Space Station — a humanoid torso that is being prepared to perform many tasks inside and outside the orbiting vehicle (1). This same kind of robotic torso has been placed on wheeled rover vehicles in a centaur- like configuration for future planetary missions, guiding the rover and picking up geological samples to analyze (1). Robots amuse children and play soccer (2), traverse the radiation-ridden earthquake-damaged nuclear reactor in Japan (3), are in development as autonomous tactical warriors for national defense (4), are planned for use in caring for the sick and elderly (5), and are operating as warehouse and industrial stock movers without human guidance (6). Most of what *can* be done with robots *will* be done, if only because there is a financial incentive for their development and use. As well, computer science is progressing at a rate that makes artificial intelligence no longer a futuristic notion: computers already outperform humans at mental challenges such as chess (7) and the trivia game “Jeopardy” (8).

A potentially negative aspect of this artificial intelligence evolution still looms, however. Scenarios envisioned by science fiction can be dismissed as the mere imaginings of artists, but such fiction sometimes contains important nuggets of possibility.

Science fiction has given us notable examples of cooperative, comforting, even entertaining robots: the archetypal “Robby the Robot” who made his first appearance in the 1956 film *Forbidden Planet*, “Data” in the *Star Trek* series, the unforgettable "C-3PO" and "R2-D2" in the *Star Wars* films. And one of the most prominent writers of the science fiction genre, Isaac Asimov, created many stories in his *Robot* series in which robots successfully served humanity by having his Three Laws of Robotics hardwired into them (9). The Laws directed that robots must primarily protect and obey human beings, and only secondarily protect themselves. Indeed, science fiction critic John Clute notes that many contemporary readers see "the very success of the *Robot* series as a demonstration that 'good' robots can exist" (10).

But there have also been darker visions, such as the guiding artificial intelligence known as "HAL 9000" in Arthur C. Clarke's *2001: A Space Odyssey*. In an interview, Stanley Kubrick (director of the film version) explained that "HAL" became a destructive agent because it experienced an "acute emotional crisis" (11). Kubrick envisioned that machines with intelligence equal to or exceeding human intelligence will "have the same emotional potentialities in their personalities as human beings" (11). The robots and intelligent machines in the film *I, Robot* were designed to follow Asimov's Three Laws, but interpreted the Laws in an unforeseen manner. They came to believe that in order to protect humanity from its self-destructive actions they must take over and exert control, sacrificing some humans in the process. The robots in the film *Blade Runner* became hazardous to humans because of their awareness of the short four-year life span humans had designed into them. The dark side of intelligent machines ran rampant in the *Matrix*, *Terminator*, and *Transformers* film series as well as the *Battlestar Galactica* television series. Even a robot specifically designed to relentlessly follow its directive to be a loving member of a human family, the robot-boy "David" in Steven Spielberg's film *A.I. Artificial Intelligence*, was portrayed as behaving in unexpected and unpredictable ways that were hazardous to humans.

Robots already possess strength beyond the abilities of the average human being; intelligence of similar magnitude is not unimaginable or impossible for robots to attain. Futurist Vernor Vinge envisions that "an ultra-intelligent machine could design even better machines", producing an accelerated evolution (12). Will such super intelligence still be controllable by humans, and still serve humanity? It is not idle speculation to think that the tables could be turned. Such scenarios are unlikely in the short- or medium-term. Yet the potential for undesired effects of the evolution of intelligent machines is nonetheless real, and we must consider this possibility and set intelligent machine evolution on a course that avoids negative trajectories.

Advances in this technological evolution may occur suddenly, as did the cloned sheep, Dolly, in the biomedical field. The journal *Science* stated with regard to Dolly: "Without prior discussion of ethical issues, the general public cannot develop a framework or common language to discuss acceptable uses of a new biomedical technology, or even whether it should be used at all" (13). A similar warning could apply to the creation of any new life and life-like forms, including synthesized biological life and intelligent machines.

Attempts to prepare for the rise of intelligent machines

A consortium of American universities produced a report called *A Roadmap for US Robotics* in 2009 (14). This 85-page document devotes considerable space to the economics and technology of robotics, with only slight mention of the possible perils of intelligent robots, recommending "the design of intrinsically safe robots with fail-safe operating systems and tools to verify the safety and correctness of robot programs" (14). This relative lack of concern may be understandable given the general perception that robots with anything like human intelligence are a long way from becoming a reality. But others, like Vernor Vinge, have a different outlook. Vinge warns: "Perhaps it was the science-fiction writers who felt the first concrete impact" of a coming change he calls a "technological singularity" (12). Comparable to the singularity of a black hole in physics, Vinge describes the technological singularity as "a point where our old models must be discarded and a new reality rules," which he views as a "change comparable to the rise of human life on Earth" (12). His vision of super intelligent entities having the ability to create even more intelligent forms depicts this evolution as potentially spiraling out of human control. Some of Vinge's predictions seem to overestimate the developmental pace of artificial intelligence. He admits that the creation of machines with intelligence greater than that of humans may be much more difficult than we now believe, and perhaps cannot be done. He notes, however, that "…if the technological singularity can happen, it will" (12).

In a post-singularity world, we may find ourselves unable to understand or even imagine the capabilities and intentions of entities with superhuman intelligence. While this future may look dim, Vinge notes: "...we are the initiators. Even the largest avalanche is triggered by small things. We have the freedom to establish initial conditions..." (12) What then, can be done at this early stage to establish the initial conditions that could prevent the dark elements of a technological singularity from coming to pass?

A model of preparation for potentially hazardous scientific and technological change

A precedent for establishing guidelines to safeguard the development of new and potentially hazardous outcomes and products can be found in the biological sciences. Our rapidly expanding knowledge of genetics has afforded us the potential power to alter these instructions and create *un*natural new life forms. While many are intrigued by the possibilities such power would bring, some have realized that these artificially created life forms could wreak havoc upon the balance of nature that evolution has established. Our own immune system, for example, could be unable to deal with artificially created viruses or bacteria that have radically new structures. Such new life forms could also invade and overwhelm established ecosystems. As these possibilities have become apparent, we have attempted to establish safeguards to deal with them.

In 1975, as DNA biotechnology was beginning to expand, concerned researchers convened a conference at Asilomar in northern California that led to the guidelines that have become the "National Institutes of Health (NIH) Guidelines for Research Involving Recombinant DNA Molecules" (15). As defined in the NIH Guidelines, recombinant DNA molecules are "...molecules that are constructed outside living cells by joining natural or synthetic DNA segments to DNA molecules that can replicate in a living cell..." (15) According to the NIH: "The purpose of the NIH Guidelines is to specify practices for constructing and handling: 1) recombinant deoxyribonucleic acid (DNA) molecules, and 2) organisms and viruses containing recombinant DNA molecules" (15). To this end the Guidelines specify "Safety Considerations", "Experiments Covered", and "Roles and Responsibilities" of the various organizations involved. The section on "Safety Considerations" consists primarily of "Risk Assessment" and "Containment", with thorough analyses of these topics by experts in the field. I suggest that exactly such analyses are also appropriate for the development of intelligent machines.

In recognition of the importance and relevance of the 1975 Asilomar Conference on biotechnology, a group of computer scientists from the Association for the Advancement of Artificial Intelligence (AAAI) convened a workshop at the same Asilomar location in February of 2009 (16). Their focus was on "whether there should be limits on research that might lead to loss of human control over computer-based systems" (17). In August 2009 this group produced the "Interim Report from the Panel Chairs", (18) detailing the nature of their concerns and the ethical and legal challenges involved. As of April 2011, their final report was not yet available, a delay possibly reflecting the lack of urgency with which the issue is presently viewed in the US.

However, much of the work in preparing for the rise of intelligent machines is being conducted outside the United States. In South Korea, there are plans to have a robot in every home by 2020, and the government is working on a "Robot Ethics Charter" (19). This document is planned to "cover standards for robotics users and manufacturers, as well as guidelines for ethical standards to be programmed into robots" (19). In Europe, the term "roboethics" has been coined for the ethics of robotics, and in January 2004 the "First International Symposium on Roboethics" (20) was held in Italy. The website www.roboethics.org reports six international conferences dealing with roboethics over the years 2004-2011 (21).

The "First International Symposium on Roboethics" in 2004 brought together humanist scholars and robotics scientists "to lay the foundations of the Ethics in the design, development and employment of the Intelligent Machines" (22). The symposium included science fiction writers, and considered such questions as "Are the intelligent robots conscious? Do they "think"? Do they feel emotions, love, pain? Once they will have learned from us everything, or understood that we are weaker than them, will they try to dominate us?" (23) The symposium also focused upon "the necessity to avoid the spread of misconceptions among the general public about the alleged dangers posed by Robotics" and focused on the creation of a "public opinion able to comprehend the positive uses of the new technology, and prevent its abuse" (23).

The symposium reviewed popular myth, legend, literature, and film (24). It was noted that it might be "possible for machines to override in-built safeguards" (25). One attendee, however, believed that "replacing biological humans with mechanical machines capable of far greater learning and cultural development is the next logical step in evolution" (25) — a view that may certainly reflect a growing conception of AI, robotics, and human-machine interaction.

The positive potential of intelligent machines

The vision of our future with intelligent machines is not universally dystopian. Inventor Ray Kurzweil[i], for instance, has written a book *The Age of Spiritual Machines* (26) in which he envisions a future where we are able to transcend all of our biological limitations, even mortality, by the creation of machines that can embody our human reality in a new nonbiological form or "substrate." This substrate would be computer hardware and software that is able to model and eventually re-create the human brain, and then evolve new, enhanced capabilities. Kurzweil thus sees the intelligent machine as part of the natural evolution of man. The 2009 film *Transcendent Man* is based upon his vision. He believes that we have evolved to create tools, and the computer is a tool that will allow us to evolve beyond our biological nature into a more hardy and long-lasting form. In this new world a human brain could be scanned and downloaded into a computer giving that "person" a new and possibly everlasting life. It may be, however, that this would prompt a redefinition of what it means to be a "person" or "human".

It might be desirable to transcend many of our biological limitations. We also strive to solve the "hard problem" of consciousness: how does the physical, objective brain produce the apparently non-physical, subjective mind and consciousness? This problem can be endlessly debated, but as noted by Daniel C. Dennett of the Institute for Cognitive Studies at Tufts University, it may be more interesting and effective to simply work on creating an artificial mind — an intelligent robot. We may then "learn something interesting about what the truly hard problems are without ever settling any of the issues about consciousness." (27)

In our attempts to create machines that can respond to sensory inputs, "think" and make decisions, and eventually become self-aware, we will be attempting to solve in step-wise fashion the problem of just what is required to produce a conscious and self-conscious being. Perhaps we will only succeed in creating machines that seem conscious but have no more real consciousness than our current computers. Even this would shed some light on the hard problem. Indeed, these are tantalizing prospects to the scientists and philosophers working in the field. Yet, we must also work to ensure that progress in resolving these philosophical problems is not accompanied by unnecessary hazards to humanity.

Specific recommendations to guide the creation of intelligent machines

One of the groups that regard the potential problems of intelligent machines seriously is the Singularity Institute for Artificial Intelligence. In a document titled "Reducing long-term catastrophic risks from artificial intelligence," (28) this group detailed not only its concerns for the risks but also its specific recommendations for dealing with them. They do not see the hazard of artificial intelligence (AI) to lie in a "robot rebellion" but in a possible competition for resources. In this scenario, AI entities with access to worldwide data networks "could radically alter their environment, e.g., by harnessing all available solar, chemical, and nuclear energy" (28) for their own purposes.

The Singularity Institute also has suggested that the current impasse in developing "superintelligent AI" is in the software, not the hardware (which is rapidly advancing). This software situation is changing, however, and it has been noted that "insights from neuroscience give advantages that past researchers lacked." Once this bottleneck is passed, progress may be sudden and AI entities could attain superhuman intelligence "more rapidly than researchers and policy-makers can develop adequate safety measures" (28).

Despite the risks, the Singularity Institute advocates that artificial intelligence is worth developing because of the benefits it will bring to humanity. Specifically, AI may allow us to break through the barriers of human ingenuity which have kept us from solving such problems as eradicating disease and averting nuclear risks. The solution for the Singularity Institute is the creation of "Friendly AI." To this end the Institute has offered software-oriented recommendations that would guarantee that an AI entity has human-friendly motivations.

Computer scientist Michael Anderson and ethicist Susan Leigh Anderson have called for programming ethical principles into robots (29), which will no doubt be important and necessary. But as science fiction so often reminds us, real-world situations can present intelligent machines with dilemmas they resolve in ways that are unexpected, unpredictable and may be hazardous to humans. We are left with the need for a fail-safe plan.

The strategy employed in the biosciences, therefore, is still appropriate. The "NIH Guidelines for Research Involving Recombinant DNA Molecules" specify containment: making sure that the spread of undesirable organisms into the environment can be stopped. Similarly, artificial intelligence research guidelines must specify emergency safeguards to stop the spread of undesirable effects from AI agents. Being able to shut off these agents is important. Few of us will accept even ethically programmed domestic robots into our homes unless they have an emergency shutoff function. The same concern is magnified on the commercial and national infrastructural levels.

Researchers at MIT have devoted considerable resources to the development of the intelligent robot "Cog." This robot has been provided with "emergency kill" mechanism(s) to provide a disabling option if its actions become harmful to humans (27). The kill button may be a simplistic concept when applied to super intelligent machines, however, since the machines themselves could understand the mechanism and perhaps override it. For this reason the continued involvement of AI programmers and developers is necessary to monitor the development of machine intelligence, and ensure that it includes a fail-safe shutoff process that could be a "sleep" rather than "kill" function, and thus not perceived as life-threatening by the machines themselves. The sleep function could also incorporate regularly scheduled time out or downtime for re-evaluation and/or reprogramming of the devices' actions. Any shutoff mechanism would require safely stopping AI operations, just as OSHA standards specify "dynamic braking systems rather than simple power cut-off" for emergency stoppage of industrial robots (30).

Conclusion

Given the economic incentives for intelligent machines, it may be impossible to arrest their development even if desired or attempted. Most of us are fascinated by intelligent machines and robots, and there are many potentially positive uses for them. It may well be that robots are a part of our own natural evolution. It is likely that intelligent robots will be developed, but on a timeline that is uncertain and unpredictable. Hence, it is important to have foundational guidelines and procedures in place to channel such development in non-hazardous directions and ensure fail-safe backup strategies. To achieve this, I believe that a government-level panel for artificial intelligence issues modeled on the Presidential Commission for the Study of Bioethical Issues, along with similar international government panels and their communication and cooperation, is essential. These groups must include input from the AI industry, scholars from the humanities, and the general public (31), and must have the power to direct and formulate guidelines for AI research, development, and use.

Disclaimer

No disclaimers.

Competing interests

The author declares he has no competing interests.

Notes

i. Creator of, among other inventions: one of the first machines able to read text aloud for blind people (first customer composer, vocalist, and musician Stevie Wonder); one of the first keyboard synthesizers able to realistically reproduce the sound of acoustic instruments (first customer also Stevie Wonder); and the Ray Kurzweil Cybernetic Poet, which is a computer able to scan the poems of a particular poet and then compose new poems in the same style (sample poems available from: http://www.kurzweilcyberart.com/poetry/rkcp_poetry_samples.php).

References

1. NASA. R2 Robonaut [Internet]. 2011. Available from: http://robonaut.jsc.nasa.gov/default.asp
2. WGBH. Soccer-Playing Robots. Nova [Internet]. 2011 Jan 27. Available from: http://www.pbs.org/wgbh/nova/tech/soccer-playing-robots.html
3. Deutsche Presse Agentur. Robots used to peer into reactor at Japan nuclear plant. The Nation [Internet]. 2011 Apr 18. Available from: http://www.nationmultimedia.com/2011/04/18/headlines/Robots-used-to-peer-into-reactor-at-Japan-nuclear-30153312.html

4. Singer P. Wired for war: the future of military robots. Brookings [Internet]. 2009 Aug 28. Available from: http://www.brookings.edu/opinions/2009/0828_robots_singer.aspx
5. Robots in Healthcare. Kinetic Consulting [Internet]. Available from: http://www.kineticconsulting.co.uk/robots.html
6. Robotic Industrial Trucks. Seegrid Corporation [Internet]. 2011. Available from: http://www.seegrid.com/
7. Deep Blue Wins. IBM [Internet]. 1997. Available from: http://www.research.ibm.com/deepblue/home/html/b.html
8. Markoff J. Computer wins on Jeopardy: trivial it's not. NY Times [Internet]. 2011 Feb 17. Available from: http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html
9. Asimov I. Run-around. New York: Street and Smith Publications; 1942
10. Clute J. Isaac Asimov. In: Seed D, editor. A companion to science fiction. Blackwell Reference Online [Internet] Blackwell Publishing; 2005. Available from: http://www.blackwellreference.com/public/tocnode?id=g9781405112185_chunk_g978140511218526
11. Gelmis J. The film director as superstar. Garden City, New York: Doubleday and Company; 1970. Excerpt Available from: http://www.visual-memory.co.uk/amk/doc/0069.html
12. Vinge V. What is the singularity? [Internet] San Diego: San Diego State University; 1993. Available from: http://mindstalk.net/vinge/vinge-sing
13. Cho MK, Magnus D, Caplan AL, McGee D. Ethical considerations in synthesizing a minimal genome. Science. 1999;286(5447):2087-90.
14. A roadmap for US robotics: From internet to robotics. [Internet] Washington, DC: Computing community consortium; 2009. Available from: http://www.us-robotics.us/reports/CCC%20Report.pdf
15. National Institutes of Health. Guidelines for research involving recombinant DNA molecules. [Internet] Available from: http://oba.od.nih.gov/oba/rac/Guidelines/NIH_Guidelines.htm
16. Association for the advancement of artificial intelligence. Presidential panel on long-term AI futures [Internet]. Available from: http://www.aaai.org/Organization/presidential-panel.php
17. Markoff J. Scientists worry machines may outsmart man. New York Times [Internet]. 2009 Jul 25. Available from: http://www.nytimes.com/2009/07/26/science/26robot.htm
18. Horvitz E, Selman B. Interim report from the panel chairs. AAAI Presidential panel on long-term AI futures [Internet]. 2009 August. Available from: http://www.aaai.org/Organization/Panel/panel-note.pdf
19. Lovgren S. Robot code of ethics to prevent android abuse, protect humans. National geographic news [Internet]. 2007 May 16. Available from: http://news.nationalgeographic.com/news/2007/03/070316-robot-ethics.html
20. Veruggio G, Operto F. Ethicsbots – Kick off meeting. Presented at Scuola di Robotica; 2005 Nov 26, Naples, Italy.
21. Roboethics [Internet]. 2004. Available from: http://www.roboethics.org
22. Veruggio G. First international symposium on roboethics: the ethical, social, humanitarian and ecological aspects of robotics [Internet]. 2004. Available from: http://www.roboethics.org/sanremo2004/
23. Veruggio G. About the symposium. First international symposium on roboethics: the ethical, social, humanitarian and ecological aspects of robotics [Internet]. 2004. Available from: http://www.roboethics.org/sanremo2004/ROBOETHICS_Symposium.html
24. Operto F. The rebellion of the machines. International symposium on roboethics: the ethical, social, humanitarian and ecological aspects of robotics [Internet]. 2004. Available from: http://www.roboethics.org/sanremo2004/ROBOETHICS_Rebellion.html
25. Operto F. Contributions. International symposium on roboethics: the ethical, social, humanitarian and ecological aspects of robotics [Internet]. 2004. Available from: http://www.roboethics.org/sanremo2004/ROBOETHICS_Contributions.html
26. Kurzweil R. The age of spiritual machines. New York: Penguin; 1999.
27. Dennett DC. Consciousness in human and robot minds. In: Ito M, Miyashita Y, Rolls ET, editors. Cognition, computation, and consciousness. Oxford: Oxford University Press; 1994. p. 17-29.
28. Reducing long-term catastrophic risks from artificial intelligence. The singularity institute for artificial intelligence [Internet]. 2011. Available from: http://singinst.org/riskintro/index.html

29. Anderson M, Anderson S. Robot be good: a call for ethical autonomous machines. Scientific American. 2010 Oct; 303 (4): 72-77.
30. OSHA. Guidelines for Robotics Safety [Internet]. 1987 Sept 21. Available from: http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=DIRECTIVES&p_id=1703
31. Giordano J, Akhouri R, McBride DK. Implantable nano-neurotechnologies: ethical, legal and social issues. J Longterm Effects Med Implants. 2009;5(9):45-54.