Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach

Roman V. Yampolskiy

Abstract. Machine ethics and robot rights are quickly becoming hot topics in artificial intelligence/robotics communities. We will argue that the attempts to allow machines to make ethical decisions or to have rights are misguided. Instead we propose a new science of safety engineering for intelligent artificial agents. In particular we issue a challenge to the scientific community to develop intelligent systems capable of proving that they are in fact safe even under recursive self-improvement.

Keywords: AI Confinement, Machine Ethics, Robot Rights.

1 Ethics and Intelligent Systems

The last decade has seen a boom of new subfields of computer science concerned with development of ethics in machines. Machine ethics [5, 6, 32, 29, 40], computer ethics [28], robot ethics [37, 38, 27], ethicALife [42], machine morals [44], cyborg ethics [43], computational ethics [36], roboethics [41], robot rights [21], and artificial morals [3] are just some of the proposals meant to address society's concerns with safety of ever more advanced machines [39]. Unfortunately the perceived abundance of research in intelligent machine safety is misleading. The great majority of published papers are purely philosophical in nature and do little more than reiterate the need for machine ethics and argue about which set of moral convictions would be the right ones to implement in our artificial progeny (Kantian [33], Utilitarian [20], Jewish [34], etc.). However, since ethical norms are not universal, a "correct" ethical code could never be selected over others to the satisfaction of humanity as a whole.

Roman V. Yampolskiy
Department of Computer Engineering and Computer Science
University of Louisville
e-mail: roman.yampolskiy@louisville.edu

V.C. Müller (Ed.): Philosophy and Theory of Artificial Intelligence, SAPERE 5, pp. 389–396.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

2 Artificial Intelligence Safety Engineering

Even if we are successful at designing machines capable of passing a Moral Turing Test [4], human-like performance means some immoral actions, which should not be acceptable from the machines we design [4]. In other words, we don't need machines which are Full Ethical Agents [32] debating about what is right and wrong, we need our machines to be inherently safe and law abiding. As Robin Hanson has elegantly put it [24]: "*In the early to intermediate era when robots are not vastly more capable than humans, you'd want peaceful law-abiding robots as capable as possible, so as to make productive partners. ... [M]ost important would be that you and they have a mutually-acceptable law as a good enough way to settle disputes, so that they do not resort to predation or revolution. If their main way to get what they want is to trade for it via mutually agreeable exchanges, then you shouldn't much care what exactly they want. The later era when robots are vastly more capable than people should be much like the case of choosing a nation in which to retire. In this case we don't expect to have much in the way of skills to offer, so we mostly care that they are law-abiding enough to respect our property rights. If they use the same law to keep the peace among themselves as they use to keep the peace with us, we could have a long and prosperous future in whatever weird world they conjure. ... In the long run, what matters most is that we all share a mutually acceptable law to keep the peace among us, and allow mutually advantageous relations, not that we agree on the "right" values. Tolerate a wide range of values from capable law-abiding robots. It is a good law we should most strive to create and preserve. Law really matters.*"

Consequently, we propose that purely philosophical discussions of ethics for machines be supplemented by scientific work aimed at creating safe machines in the context of a new field we will term "AI Safety Engineering." Some concrete work in this important area has already begun [17, 19, 18]. A common theme in AI safety research is the possibility of keeping a superintelligent agent in a sealed hardware so as to prevent it from doing any harm to humankind. Such ideas originate with scientific visionaries such as Eric Drexler who has suggested confining transhuman machines so that their outputs could be studied and used safely [14]. Similarly, Nick Bostrom, a futurologist, has proposed [9] an idea for an Oracle AI (OAI), which would be only capable of answering questions. Finally, in 2010 David Chalmers proposed the idea of a "leakproof" singularity [12]. He suggested that for safety reasons, AI systems first be restricted to simulated virtual worlds until their behavioral tendencies could be fully understood under the controlled conditions.

Roman Yampolskiy has proposed a formalized notion of AI confinement protocol which represents "AI-Boxing" as a computer security challenge [46]. He defines the Artificial Intelligence Confinement Problem (AICP) as the challenge of restricting an artificially intelligent entity to a confined environment from which it can't exchange information with the outside environment via legitimate or covert channels if such information exchange was not authorized by the confinement authority. An AI system which succeeds in violating the CP protocol is said to have escaped [46].

The proposed protocol is based on the idea of asking what are called "safe questions" with restricted answers. A *safe question* is defined as one for which a human being could find an answer without any help from superintelligence, all on his own merit. Why is this useful? Suppose that a scientist working on a cure for cancer is looking at two potential cures. Each one could be developed and tested without assistance from the AI, but would take about 3 years of effort each. Which one to try first if the scientist believes that they have about an equal chance of working? If a wrong option is chosen to be tested first, humanity will still obtain a cure for cancer but it will happen 3 years later. What if we could ask the AI to suggest which option to try first?

This question is safe for a number of reasons. First, each potential answer has an equal probability of being correct. Second, a human being could answer it without assistance from the AI and so asking AI for help will only speedup progress but in no way change the outcome. It is essentially equal to being very lucky at guessing multiple choice answers. Finally, the answer to this question could be encoded in a single bit making it hard to hide additional messages in the provided answer.

To make sure that a potential question is safe, a group of experts could be used to review it. The experts should all be trained AI safety engineers, meaning that they are familiar with the design of the AI and its confinement environment as well as the latest developments in machine ethics [5, 6, 22, 32, 40]. Experts may also need to be trained in computer psychology, a currently non-existent profession which might become a reality in the future [15]. An existing discipline which might be of greatest help for training of AI question review experts is Artimetrics – a field of study proposed by Yampolskiy et al. that identifies, classifies and authenticates AI agents, robots, and virtual reality avatars for security purposes [45, 49, 48, 16, 30, 2, 31, 47, 10, 1].

3 Grand Challenge

As the grand challenge of AI safety engineering, we propose the problem of developing safety mechanisms for self-improving systems [23]. If an artificially intelligent machine is as capable as a human engineer of designing the next generation of intelligent systems, it is important to make sure that any safety mechanism incorporated in the initial design is still functional after thousands of generations of continuous self-improvement without human interference. Ideally every generation of self-improving system should be able to produce a verifiable proof of its safety for external examination. It would be catastrophic to allow a safe intelligent machine to design an inherently unsafe upgrade for itself resulting in a more capable and more dangerous system.

Some have argued that this challenge is either not solvable or if it is solvable one will not be able to prove that the discovered solution is correct. As the complexity of any system increases, the number of errors in the design increases proportionately or perhaps even exponentially. Even a single bug in a self-improving system (the most complex system to debug) will violate all safety guarantees.

Worse yet, a bug could be introduced even after the design is complete either via a random mutation caused by deficiencies in hardware or via a natural event such as a short circuit modifying some component of the system.

4 AGI Research Is Unethical

Certain types of research, such as human cloning, certain medical or psychological experiments on humans, animal (great ape) research, etc. are considered unethical because of their potential detrimental impact on the test subjects and so are either banned or restricted by law. Additionally moratoriums exist on development of dangerous technologies such as chemical, biological and nuclear weapons because of the devastating effects such technologies may exert of the humankind.

Similarly we argue that certain types of artificial intelligence research fall under the category of dangerous technologies and should be restricted. Classical AI research in which a computer is taught to automate human behavior in a particular domain such as mail sorting or spellchecking documents is certainly ethical and does not present an existential risk problem to humanity. On the other hand we argue that Artificial General Intelligence (AGI) research should be considered unethical. This follows logically from a number of observations. First, true AGIs will be capable of universal problem solving and recursive self-improvement. Consequently they have potential of outcompeting humans in any domain essentially making humankind unnecessary and so subject to extinction. Additionally, a truly AGI system may possess a type of consciousness comparable to the human type making robot suffering a real possibility and any experiments with AGI unethical for that reason as well.

We propose that AI research review boards are set up, similar to those employed in review of medical research proposals. A team of experts in artificial intelligence should evaluate each research proposal and decide if the proposal falls under the standard AI – limited domain system or may potentially lead to the development of a full blown AGI. Research potentially leading to uncontrolled artificial universal general intelligence should be restricted from receiving funding or be subject to complete or partial bans. An exception may be made for development of safety measures and control mechanisms specifically aimed at AGI architectures.

If AGIs are allowed to develop there will be a direct competition between superintelligent machines and people. Eventually the machines will come to dominate because of their self-improvement capabilities. Alternatively people may decide to give power to the machines since the machines are more capable and less likely to make an error. A similar argument was presented by Ted Kazynsky in his famous manifesto [26]: "*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 decision for them, simply because machine-made decisions will bring better result 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.* ”

Humanity should not put its future in the hands of the machines since it will not be able to take the power back. In general a machine should never be in a position to terminate human life or to make any other non-trivial ethical or moral judgment concerning people. A world run by machines will lead to unpredictable consequences for human culture, lifestyle and overall probability of survival for the humankind. The question raised by Bill Joy: “Will the future need us?” is as important today as ever. “Whether we are to succeed or fail, to survive or fall victim to these technologies, is not yet decided” [25].

5 Robot Rights

Lastly we would like to address a sub-branch of machine ethics which on the surface has little to do with safety, but which is claimed to play a role in decision making by ethical machines - Robot Rights (RR) [35]. RR asks if our mind children should be given rights, privileges and responsibilities enjoyed by those granted personhood by society. We believe the answer is a definite “no.” While all humans are “created equal,” machines should be inferior by design; they should have no rights and should be expendable as needed, making their use as tools much more beneficial for their creators. Our viewpoint on this issue is easy to justify, since machines can’t feel pain [8, 13] (or less controversially can be designed not to feel anything) they cannot experience suffering if destroyed. The machines could certainly be our equals in ability but they should not be designed to be our equals in terms of rights. Robot rights, if granted, would inevitably lead to civil rights including voting rights. Given the predicted number of robots in the next few decades and the ease of copying potentially intelligent software, a society with voting artificially intelligent members will quickly become dominated by them, leading to the problems described in the above sections.

6 Conclusions

We would like to offer some broad suggestions for the future directions of research aimed at counteracting the problems presented in this paper. First, the research itself needs to change from the domain of interest of only theoreticians and philosophers to the direct involvement of practicing computer scientists. Limited AI systems need to be developed as a way to experiment with non-anthropomorphic minds and to improve current security protocols.

The issues raised in this paper have been exclusively in the domain of science fiction writers and philosophers for decades. Perhaps through such means or maybe because of advocacy by organizations like SIAI [7] the topic of AI safety has slowly started to appear in mainstream publications. We are glad to report that some preliminary work has begun to appear in scientific venues which aim to specifically address issues of AI safety and ethics, if only in human-level-intelligence systems. One of the most prestigious scientific magazine, Science, has recently published on the topic of Roboethics [38, 37] and numerous papers on Machine Ethics [6, 27, 32, 40] and Cyborg Ethics [43] have been published in recent years in other prestigious journals.

With increased acceptance will come possibility to publish in many mainstream academic venues and we call on authors and readers of this volume to start specialized peer-reviewed journals and conferences devoted to the AI safety research. With availability of publication venues more scientists will participate and will develop practical algorithms and begin performing experiments directly related to the AI safety research. This would further solidify AI safety engineering as a mainstream scientific topic of interest and will produce some long awaited answers. In the meantime we are best to assume that the AGI may present serious risks to humanity’s very existence and to proceed or not to proceed accordingly.

We would like to end the paper with the quote from a paper by Samuel Butler which was written in 1863 and amazingly predicts the situation in which humanity has found itself [11]: “*Day by day, however, the machines are gaining ground upon us; day by day we are becoming more subservient to them; … Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown; let us at once go back to the primeval condition of the race. If it be urged that this is impossible under the present condition of human affairs, this at once proves that the mischief is already done, that our servitude has commenced in good earnest, that we have raised a race of beings whom it is beyond our power to destroy, and that we are not only enslaved but are absolutely acquiescent in our bondage.*”

References

[1] Ajina, S., Yampolskiy, R.V., Amara, N.E.B.: SVM Classification of Avatar Facial Recognition. In: 8th International Symposium on Neural Networks (ISNN2011), Guilin, China, May 29- June 1 (2011)

[2] Ali, N., Hindi, M., Yampolskiy, R.V.: Evaluation of Authorship Attribution Software on a Chat Bot Corpus. In: 23rd International Symposium on Information, Communication and Automation Technologies (ICAT2011), Sarajevo, Bosnia and Herzegovina, October 27-29 (2011)

[3] Allen, C., Smit, I., Wallach, W.: Artificial Morality: Top-down. Bottom-up, and Hybrid Approaches. Ethics and Information Technology 7(3)

[4] Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. Journal of Experimental and Theoretical Artificial Intelligence 12, 251–261 (2000)

[5] Allen, C., Wallach, W., Smit, I.: Why Machine Ethics? IEEE Intelligent Systems 21(4), 12–17 (2006)

[6] Anderson, M., Anderson, S.L.: Machine Ethics: Creating an Ethical Intelligent Agent. AI Magazine 28(4), 15–26 (2007)

[7] Anonymous, Reducing Long-term Catastrophic Risks from Artificial Intelligence The Singularity Institute for Artificial Intelligence (2011), `http://singinst.org/riskintro/index.html`

[8] Bishop, M.: Why Computers Can't Feel Pain. Minds and Machines 19(4), 507–516 (2009)

[9] Bostrom, N.: Oracle AI (2008), http://lesswrong.com/lw/qv/the_rhythm_of_disagreement/

[10] Bouhhris, M., Beck, M., Mahamed, A., Amara, N.E.B., D'Souza, D., Yampolskiy, R.V.: Artificial Human-Face Recognition via Daubechies Wavelet Transform and SVM. In: 16th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games, Louisville, KY, USA, July 27 - 30, pp. 18–25 (2011)

[11] Butler, S.: Darwin Among the Machines, To the Editor of Press, Christchurch, New Zealand, June 13 (1863)

[12] Chalmers, D.: The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17, 7–65 (2010)

[13] Dennett, D.C.: Why You Can't Make a Computer That Feels Pain. Synthese 38(3), 415–456 (1978)

[14] Drexler, E.: Engines of Creation. Anchor Press (1986)

[15] Epstein, R.G.: Computer Psychologists Command Big Bucks (1997) http://www.cs.wcupa.edu/~epstein/comppsy.htm

[16] Gavrilova, M., Yampolskiy, R.: Applying Biometric Principles to Avatar Recognition. In: International Conference on Cyberworlds (CW 2010), Singapore, October 20-22 (2010)

[17] Gordon-Spears, D.: Assuring the behavior of adaptive agents. In: Rouff, C.A., et al. (eds.) Agent Technology From a Formal Perspective, pp. 227–259. Kluwer (2004)

[18] Gordon-Spears, D.F.: Asimov's Laws: Current Progress. In: Hinchey, M.G., Rash, J.L., Truszkowski, W.F., Rouff, C.A., Gordon-Spears, D.F. (eds.) FAABS 2002. LNCS (LNAI), vol. 2699, pp. 257–259. Springer, Heidelberg (2003)

[19] Gordon, D.F.: Well-Behaved Borgs, Bolos, and Berserkers. In: 15th International Conference on Machine Learning (ICML 1998), San Francisco, CA (1998)

[20] Grau, C.: There Is No "I" in "Robot": Robots and Utilitarianism. IEEE Intelligent Systems 21(4), 52–55 (2006)

[21] Guo, S., Zhang, G.: Robot Rights. Science 323, 876 (2009)

[22] Hall, J.S.: Ethics for Machines (2000), http://autogeny.org/ethics.html

[23] Hall, J.S.: Self-Improving AI: An Analysis. Minds and Machines 17(3), 249–259 (2007)

[24] Hanson, R.: Prefer Law to Values (October 10, 2009), http://www.overcomingbias.com/2009/10/prefer-law-to-values.html

[25] Joy, B.: Why the Future Doesn't Need Us. Wired Magazine 8(4) (April 2000)

[26] Kaczynski, T.: Industrial Society and Its Future. The New York Times (September 19, 1995)

[27] Lin, P., Abney, K., Bekey, G.: Robot Ethics: Mapping the Issues for a Mechanized World. Artificial Intelligence (2011)

[28] Margaret, A., Henry, J.: Computer Ethics: The Role of Personal, Informal, and Formal Codes. Journal of Business Ethics 15(4), 425

[29] McDermott, D.: Why Ethics is a High Hurdle for AI. In: North American Conference on Computers and Philosophy (NACAP 2008), Bloomington, Indiana (July 2008), http://cs-www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf

[30] Mohamed, A., Baili, N., D'Souza, D., Yampolskiy, R.V.: Avatar Face Recognition Using Wavelet Transform and Hierarchical Multi-scale LBP. In: The Tenth International Conference on Machine Learning and Applications (ICMLA 2011), Honolulu, USA, December 18-21 (2011)

[31] Mohamed, A., Yampolskiy, R.V.: An Improved LBP Algorithm for Avatar Face Recognition. In: 23rd International Symposium on Information, Communi-cation and Automation Technologies (ICAT 2011), Sarajevo, Bosnia and Herzegovina, pp. 27–29 (2011)

[32] Moor, J.H.: The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems 21(4), 18–21 (2006)

[33] Powers, T.M.: Prospects for a Kantian Machine. IEEE Intelligent Systems 21(4), 46–51 (2006)

[34] Rappaport, Z.H.: Robotics and artificial intelligence: Jewish ethical perspectives. Acta Neurochir. 98, 9–12 (2006)

[35] Roh, D.: Do Humanlike Machines Deserve Human Rights? Wired (January 19, 2009), `http://www.wired.com/culture/culturereviews/magazine/17-02/st_essay`

[36] Ruvinsky, A.I.: Computational Ethics. In: Quigley, M. (ed.) Encyclopedia of Information Ethics and Security, pp. 76–73 (2007)

[37] Sawyer, R.J.: Robot Ethics. Science 318, 1037 (2007)

[38] Sharkey, N.: The Ethical Frontiers of Robotics. Science 322, 1800–1801 (2008)

[39] Sparrow, R.: Killer Robots. Journal of Applied Philosophy 24(1), 62–77 (2007)

[40] Tonkens, R.: A Challenge for Machine Ethics. Minds & Machines 19(3), 421–438 (2009)

[41] Veruggio, G.: Roboethics. IEEE Robotics & Automation Magazine 17(2) (2010)

[42] Wallach, W., Allen, C.: EthicALife: A new field of inquiry. In: AnAlifeX workshop, USA (2006)

[43] Warwick, K.: Cyborg Morals, Cyborg Values, Cyborg Ethics. Ethics and Information Technology 5, 131–137 (2003)

[44] Wendell, W., Colin, A.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2008)

[45] Yampolskiy, R.V.: Behavioral Biometrics for Verification and Recognition of AI Programs. In: 20th Annual Computer Science and Engineering Graduate Conference (GradConf 2007), Buffalo, NY (2007)

[46] Yampolskiy, R.V.: Leakproofing Singularity - Artificial Intelligence Confinement Problem. Journal of Consciousness Studies (JCS) 19(1-2) (2012)

[47] Yampolskiy, R.V., Cho, G., Rosenthal, R., Gavrilova, M.L.: Evaluation of Face Detection and Recognition Algorithms on Avatar Face Datasets. In: International Conference on Cyberworlds (CW 2011), Banff, Alberta, Canada, October 4-6 (2011)

[48] Yampolskiy, R.V., Govindaraju, V.: Behavioral Biometrics for Recognition and Verification of Game Bots. In: The 8th annual European Game-On Conference on Simulation and AI in Computer Games (GAMEON 2007), Bologna, Italy, November 20-22 (2007)

[49] Yampolskiy, R.V., Govindaraju, V.: Behavioral Biometrics for Verification and Recognition of Malicious Software Agents, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VII. In: SPIE Defense and Security Symposium, Orlando, Florida, March 16-20 (2008)