Attempts to Attribute Moral Agency to Intelligent Machines are Misguided

Roman V. Yampolskiy
University of Louisville
roman.yampolskiy@louisville.edu

Abstract

Machine ethics are quickly becoming an important part of artificial intelligence research. We argue that attempts to attribute moral agency to intelligent machines are misguided, whether applied to infrahuman or superhuman AIs. Humanity should not put its future in the hands of the machines that do not do exactly what we want them to, since we will not be able to take power back. In general, a machine should never be in a position to make any non-trivial ethical or moral judgments concerning people unless we are confident, preferably with mathematical certainty, that these judgments are what we truly consider ethical.

1 Ethics and Intelligent Systems

The last decade has seen a boom in the field of computer science concerned with the application of ethics to machines that have some degree of autonomy in their action. Variants under names such as machine ethics (Allen, Wallach, & Smit, 2006; Anderson & Anderson, 2007; Hall, 2007a; McDermott, 2008; Moor, 2006; Tonkens, 2009) computer ethics (Pierce & Henry, 1996), robot ethics (Lin, Abney, & Bekey, 2011; Sawyer, 2007; Sharkey, 2008), ethicALife (Wallach & Allen, 2006), machine morals (Wallach & Allen, 2008), cyborg ethics (Warwick, 2003), computational ethics (Ruvinsky, 2007), roboethics (Veruggio, 2010), robot rights (Guo & Zhang, 2009), artificial morals (Allen, Smit, & Wallach, 2005), and Friendly AI (Yudkowsky, 2008) are some of the proposals meant to address society's concerns with the ethical and safety implications of ever more advanced machines (Sparrow, 2007).

Unfortunately, the rapid growth of research in intelligent-machine ethics and safety has not brought real progress. The great majority of published papers do little more than argue about which of the existing schools of ethics, built over the centuries to answer the needs of a human society, would be the right one to implement in our artificial progeny: Kantian (Powers, 2006), deontological (Anderson & Anderson, 2007; Asimov, 1942), utilitarian (Grau, 2006), Jewish (Rappaport, 2006), and others.

Moreover, machine ethics discusses machines with roughly human-level intelligence or below, not machines with far-above-human intelligence (Yampolskiy, 2013). Yet the differences between infrahuman, human-level, and superhuman intelligences are essential (Hall, 2007a, 2007b). We generally do not ascribe moral agency to infrahuman agents such as non-human animals. Indeed, even humans with less than full intelligence, like children and those with severe intellectual disability, are excluded from moral agency, though still considered moral patients, the objects of responsibility for moral agents. All existing AIs are infrahuman when judged in terms of flexible, general intelligence. Human-level AIs, if similar to humans in their mental goals and architecture, should be treated by the same ethical considerations applied to humans, but if they are deeply inhuman in their mental

architecture, some of the usual considerations may fail. In this article, we will consider safety factors for AIs at a roughly human level of ability or above, referred to by the new term of art "artificial general intelligence."[*]

2 Ethics of Superintelligence

Even more important than infrahuman and near-human AIs are superintelligent AIs. A roughly human-level machine is likely to soon become superhuman, so that the latter are more likely to be widespread in our future than near-human AIs (Chalmers, 2010). Once an AI is developed with roughly human levels of ability, it will seek the best techniques for achieving its aims. One useful technique is to improve intelligence in itself or in a new generation of AIs (Omohundro, 2008). If, based on general-purpose computer infrastructure, an AI will be able to add hardware; it will also be able to improve its software by continuing the work that the human engineers used to bring it up to its present level.

The human level of intelligence has prominence as the level available to our observation. It happens to be the lowest level capable of forming a civilization—no life form with lower intelligence has done so to date, but humans have. It also seems to be, if predictions about coming decades come true, the lowest level capable of engineering a new type of intelligence. Yet physical laws allow far higher levels of processing power, and probably of intelligence (Sotala, 2010). These levels can be reached with recursive self-improvement. In the words of IJ Good (1965):

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind.

Such a machine may surpass humans in "all the intellectual activities of any man," or just in some of them; it may have intellectual capacities that no human has. If today's trends continue, by 2049, $1000 will buy computer power exceeding the computational capacities of the entire human species (Kurzweil, 2006). If true artificial general intelligence is established and can take full advantage of such raw power, it will have advantages not shared by humans. Human computational capacity does not rise linearly in effectiveness as people are added, whereas computers might be able to make greater use of their computational power. Computers can introspect, self-improve, and avoid biases imposed by ancestral heuristics, among other human limitations (Sotala, 2012).

More important than the exact areas in which the agent is specialized is the effect that it can have on people and their world, particularly if it is much more powerful than humans. For this reason, we should understand intelligence abstractly and generally as the ability to achieve complex goals in complex environments (Legg & Hutter, 2007) rather than on the human model. A vastly superhuman intelligence could have extreme effects on all humanity: Indeed, humans today have the power to destroy much of humanity with nuclear weapons, and a fortiori a superhuman intelligence could do so. A superintelligence, if it were so powerful that humans could not have meaningful effect on the achievement of its goals, would not be constrained by promises and threats of rewards and punishment, as humans are. The human brain architecture and goal systems, including ethical mental systems, are complex function-specific structures contingent on the environments in which the human

[*] The term AGI can also refer more narrowly to engineered AI, in contrast to those derived from the human model, such as emulated or uploaded brains. (Goertzel & Pennachin, 2007). In this article, unless specified otherwise, we use AI and AGI to refer to artificial general intelligences in the broader sense.

species developed (Churchland, 2011; Tooby & Cosmides, 1992; Wright, 2001). Most possible mind architectures and goal systems are profoundly non-anthropomorphic (where "anthrophomorphic," for our purposes, means "a mind having human-like qualities"). Only if it is specifically based on the human model will a newly created mind resemble ours (Yampolskiy & Fox, 2012a) (Muehlhauser and Helm, "The Singularity and Machine Ethics," in press). Thus, future AIs pose very different ethical questions from human agents.

Defining an ethical system for a superhuman and inhuman intelligence takes us to areas inadequately explored by philosophers to date. Any answer must be based on common human ethical values rooted in our shared history. These are a complex and inconsistent mixture, similar but not identical across societies and among individuals. Despite many areas of commonality, ethical norms are not universal, and so a single "correct" deontological code based on any predefined abstract principles could never be selected over others to the satisfaction of humanity as a whole; nor could the moral values of a single person or culture be chosen for all humanity.

Asimov's (1942) Laws of Robotics are often cited as a deontological approach to ethical robot behavior and have inspired numerous imitations as well as critique (Gordon-Spears, 2003; LaChat, 1986; McCauley, 2007; Pynadath & Tambe, 2001; Weld & Etzioni, 1994). The original laws as given by Asimov are (Asimov, March 1942):

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with either the First or Second Law.

Clarke (Clarke, 1993, 1994), arguably, provides the best analysis of implications of Asimov's work on information technology. In particular he brings up the issues of linguistic ambiguity, the role of judgment in decision making, conflicting orders, valuation of humans, and many others. It must be emphasized that Asimov wrote fiction. His writing was optimized for an interesting and plausible plot, not for accurate prediction. The "good story bias" (Bostrom, 2002) towards scenarios that make a good plot, like laws of robot ethics that *fail* in each story, is useful in fiction, but dangerous in speculation about real life. Even to the extent that the plots in Asimov's stories are plausible, they and others like them represent only a few scenarios from a much broader space of possibilities. It would be a mistake to focus on the narrow examples that have been described in fiction, rather than to try to understand the full range of possibilities ahead of us (Yudkowsky, 2007). The general consensus seems to be that no set of rules can ever capture every possible situation and that interaction of rules may lead to unforeseen circumstances and undetectable loopholes leading to devastating consequences for the humanity (Yampolskiy, October 3-4, 2011).

Whatever the rules imposed, it would be dangerous to attempt to constrain the behavior of advanced artificial intelligences which interpret these rules without regard for the complex ensemble of human values. Simple constraints on behavior have no value when AIs which are smarter than humans and so can bypass these rules, if they so choose. They may take their behavior in dangerous new directions when facing challenges and environments never before seen by human beings, and not part of the set of situations used to program, train, or test their behavior (Yudkowsky, 2008; Yudkowsky & Bostrom, 2011).

Even if we are successful at designing machines capable of passing a Moral Turing Test (Allen, Varner, & Zinser, 2000), that is, those that can successfully predict humans' answers on moral questions, we would not have created the ultimate moral machines. Such tests test prediction power, not motivation to act on moral principles. Moreover, emulating humans is not moral perfection: Humans err in moral questions, even according to their own judgment, and we should preferably

avoid such imperfection in machines we design (Allen et al., 2000). This is all the more true for machines more powerful than us.

We do not want our machine-creations behaving in the same way humans do (Fox, 2011). For example, we should not develop machines which have their own survival and resource consumption as terminal values, as this would be dangerous if it came into conflict with human well-being. Likewise, we do not need machines that are Full Ethical Agents (Moor, 2006) deliberating about what is right and coming to uncertain solutions; we need our machines to be inherently stable and safe. Preferably, this safety should be mathematically provable.

At an early stage, when AIs have near-human intelligence, and perhaps humanlike mind architectures and motivation systems, humanlike morality, regulated by law, trade, and other familiar constraints towards mutual cooperation, may be enough.

In the words of Robin Hanson (2010):

> 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.

Hanson extrapolates this dynamic to a later world with superhuman minds:

> [In t]he later era when robots are vastly more capable than people… 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.

This extrapolation is incorrect, at least if those minds are non-anthropomorphic. Such law-abiding tendencies cannot be assumed in superintelligences (Fox & Shulman, 2010). Direct instrumental motivations—the fear of punishment and desire for the benefits of cooperation—will not function for them. An AI far more powerful than humans could evade monitoring and resist punishment. It would have no need for any benefits that humans could offer in exchange for its good behavior. The Leviathan state (Hobbes, 1998/1651), enforcing mutual cooperation through laws, has no inherent significance if a single intelligence is far more powerful than the entire state. Thus, direct reward and punishment will not be sufficient to cause all superhuman AIs to cooperate.

Going beyond simple reciprocity, trustworthy benevolent dispositions can also serve to ensure instrumental cooperation. If one can reliably signal trustworthiness to others, then one's disposition can engender trust and so increase mutual cooperation, even in cases where breaking the trust would provide net benefit (Gauthier, 1986).

An AI built in the Artificial General Intelligence paradigm, in which the design is engineered de novo, has the advantage over humans with respect to transparency of disposition, since it is able to display its source code, which can then be reviewed for trustworthiness (Salamon, Rayhawk, & Kramár, 2010; Sotala, 2012). Indeed, with an improved intelligence, it might find a way to formally prove its benevolence. If weak early AIs are incentivized to adopt verifiably or even provably benevolent dispositions, these can be continually verified or proved and thus retained, even as the AIs gain in intelligence and eventually reach the point where they have the power to renege without retaliation (Hall, 2007a).

Nonetheless, verifiably benevolent dispositions would not necessarily constrain a superintelligence AI. If it could successfully signal a benevolent disposition that it does not have, it can do even better. If its ability to deceive outpaces its ability to project signals of benevolence verifiable by humans, then the appearance of a benevolent disposition would do more harm than good.

We might hope that increased intelligence would lead to moral behavior in an AI by structuring terminal values. Chalmers (2010) asks whether a superintelligence would necessarily have morality as an end-goal. Yet theoretical models such as AIXI (Hutter, 2005) specify systems with maximal intelligence, across all possible reward functions. There is no reason that a superintelligence would necessarily have goals favoring human welfare, which are a tiny part of the space of possible goals.

Nor can we assume that a superintelligence would undergo a Kantian shift towards a moral value system. If a system is working towards a given goal, then changes to that goal make it less likely that the goal will be achieved. Thus, unless it had higher-order terminal values in favor of goal-changing, it would do whatever is necessary to protect its goals from change (Omohundro, 2008).

> Consider Gandhi, who seems to have possessed a sincere desire not to kill people. Gandhi would not knowingly take a pill that caused him to want to kill people, because Gandhi knows that if he wants to kill people, he will probably kill people, and the current version of Gandhi does not want to kill (Yudkowsky & Bostrom, 2011)

An intelligence will consume all possible resources in achieving its goals, unless its goals specify otherwise. If a superintelligence does not have terminal values that specifically optimize for human well-being, then it will compete for resources that humans need, and since it is, by hypothesis, much more powerful than humans, it will succeed in monopolizing all resources. To survive and thrive, humans require mass and energy in various forms, and these can be expected to also serve for the achievement of the AI's goals. We should prevent the development of an agent that is more powerful than humans are and that competes over such resources.

3 Unconstrained AI Research is Unethical

Some types of research, such as certain medical or psychological experiments on humans, are considered potentially unethical because of the possibility of detrimental impact on the test subjects, treated as moral patients; such research is thus either banned or restricted by law. Experiments on animals have also been restricted. Additionally, moratoriums exist on development of dangerous technologies such as chemical, biological, and nuclear weapons because of the devastating effects such technologies may have on humanity.

Since the 1970s, institutional review boards have overseen university research programs in the social and medical sciences; despite criticism and limited formal enforcement power, these boards have proven able to regulate experimental practices.

In the sphere of biotechnology, the Asilomar Conference on Recombinant DNA drew up rules to limit the cross-species spread of recombinant DNA by defining safety standards, for example containing biohazards in laboratories. The guidelines also prohibited certain dangerous experiments like the cloning of pathogens (Berg, Baltimore, Brenner, Roblin, & Singer, 1975). Despite the temptation for scientists to gain a competitive edge by violating the principles, the scientific community has largely adhered to these guidelines in the decades since.

Similarly, we argue that certain types of artificial intelligence research fall under the category of dangerous technologies, and should be restricted. Narrow AI research, for example in the automation of human behavior in a specific domain such as mail sorting or spellchecking, is certainly ethical, and does not present an existential risk to humanity. On the other hand, research into artificial general intelligence, without careful safety design in advance, is unethical. Since true AGIs will be capable of universal problem solving and recursive self-improvement, they have the potential to outcompete humans in any domain. Humans are in danger of extinction if our most basic resources are lost to AIs outcompeting us.

In addition, depending on its design, and particularly if it is modeled after the human example, a flexible and general artificial intelligence may possess those aspects of the human mind that grant moral patient status—for example, the capacity to feel physical or mental pain—making robot suffering a real possibility, and rendering unethical a variety of experiments on the AI.

We propose that AI research review boards be set up, comparable to those employed in the review of medical research proposals. A team of experts in artificial intelligence, with training in the novel ethical questions posed by advanced AI, should evaluate each research proposal and decide if it falls under the category of narrow AI, or if it may potentially lead to the development of a full, flexible, AGI. The latter should be restricted with appropriate measures, ranging from supervision, to funding limits, to a partial or complete ban. At the same time, research focusing on the development of safety measures for AGI architectures should be encouraged, as long as that research does not pose risks incommensurate with the potential benefits.

If AIs at human level and above are developed, the human species will be at risk, unless the machines are specifically designed to pursue human welfare, correctly defined, as their primary goal. Machines not designed for such "Friendliness," to use the technical term of art, will come to destroy humanity as a side effect of its goal-seeking, since resources useful to humanity will likely also be found useful by a superintelligence. The alternative is to define the correct goal system and mechanism for preserving it, and then reap the benefits of this superintelligent instrument of the human will.

The risk from superintelligence machines is extinction, not domination. Some fear the latter, as in the manifesto of Ted Kaczynski (1995)

> 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.

Kaczynski, who gained his fame as the Unabomber through a terror campaign, makes an assumption that calls into question the implicit conclusion of this quote. The words "hand over all the power" and "the machines will be in… control" assume that the machines will be in an adversarial position; that they will seek to dominate humanity for purposes of their own. But the desire for domination of the other is a characteristic of humans and other animals, which developed because of its adaptive value.

Domination of humans would indeed be useful to an AI whose goals did not treat human values as primary, so long as the AI remains at near-human levels. Yet at superintelligent levels, the analogy to human tyranny fails. If, on the one hand, superintelligent machines have goals that do not correspond to human values, the likely result is human extinction. Intelligent agents who are many orders of magnitude more capable than humans will be able to achieve goals without the help of humans, and will most likely use up resources essential to human survival in doing so. (An exception would be if the machines have human enslavement as a terminal value in its own right.) On the other hand, superintelligent machines whose goal is to allow humans to achieve their values will work effectively to maximize for those values. Freedom is one such value, and so would also be part of the AI's goal-system, subject to the need to preserve other human values. If such human-friendly AIs do come into being, they will indeed have tremendous power in shaping the world, but they will still be tools for the

benefit of humanity. We humans now depend on technology such as modern farming, transportation, and public-health systems. If these were removed, the human future would be at risk, yet we generally do not fear these technologies, because they exist to serve us. So too would super-powerful intelligent agents serve as worthy tools, so long as their goal system is correctly defined.

Still, we should take this precaution: Humanity should not put its future in the hands of the machines that do not do exactly what we want them to, since we will not be able to take power back. In general, a machine should never be in a position to make any non-trivial ethical or moral judgments concerning people unless we are confident, preferably with mathematical certainty, that these judgments are what we truly consider ethical. A world run by machines whose goal systems were not precisely tuned to our needs would lead to unpredictable, and probably extremely dangerous, consequences for human culture, lifestyle, and survival. The question raised by Bill Joy (2000), "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."

4 Conclusions

We would like to offer some suggestions for the possible directions of future research aimed at addressing the problems presented above. First, as the implications of future artificial general intelligence become clearer, and even before artificial general intelligence is actually implemented, progress in several new research areas must grow rapidly. Theoretical and practical research into AI safety needs to be ramped up significantly, with the direct involvement of decision theorists, neuroscientists, and computer scientists, among other specialists. Limited AI systems need to be developed to allow direct experimentation with infrahuman minds, but in all cases with a careful consideration of risks and security protocols (Yampolskiy, 2011).

Work in infrahuman and human-level AI ethics is becoming more common, and has begun to appear in scientific venues that aim to specifically address issues of AI safety and ethics. The journal *Science* has recently published on the topic of roboethics (Sawyer, 2007; Sharkey, 2008), and numerous papers on machine ethics (Anderson & Anderson, 2007; Lin et al., 2011; Moor, 2006; Tonkens, 2009) and cyborg ethics (Warwick, 2003) have been published in recent years in other prestigious journals. Most such writing focuses on infrahuman systems, avoiding the far more interesting and significant implications of human-level and superintelligent AI.

We call on researchers to advance this research to encompass superintelligent machines and to concentrate on developing safe artificial general intelligence. The work should focus on safety mechanisms, while also supporting the growth of a field of research with important theoretical and practical implications. Humanity needs the theory, the algorithms, and eventually the implementation of rigorous safety mechanisms, starting in the very first AI systems. In the meantime, we should assume that AGI may present serious risks to humanity's very existence, and carefully restrain our research directions accordingly.

Acknowledgements

Ideas presented in this work have been previously discussed in "Artificial intelligence safety engineering: Why machine ethics is a wrong approach" (Yampolskiy, 2011) and in "Safety Engineering for Artificial General Intelligence" (Yampolskiy & Fox, 2012b).

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3).

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence, 12*, 251-261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems, 21(4)*, 12-17.

Anderson, M., & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine, 28(4)*, 15-26.

Asimov, I. (1942, March). Runaround. *Astounding Science Fiction*, 94-103.

Asimov, I. (March 1942). *Runaround in Astounding Science Fiction*.

Berg, P., Baltimore, D., Brenner, S., Roblin, R. O., & Singer, M. F. (1975). Summary statement of the Asilomar Conference on Recombinant DNA Molecules. *Proceedings of the National Academy of Sciences of the United States of America, 72*(6), 1981-1984.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*(1).

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*, 7-65.

Churchland, P. S. (2011). *Brain trust*. Princeton, NJ: Princeton University Press.

Clarke, R. (1993). Asimov's Laws of Robotics: Implications for Information Technology, Part 1. *IEEE Computer, 26(12)*, 53-61.

Clarke, R. (1994). Asimov's Laws of Robotics: Implications for Information Technology, Part 2. *IEEE Computer, 27(1)*, 57-66.

Fox, J. (2011). *Morality and super-optimizers*. Paper presented at the Future of Humanity Conference, Oct. 24, 2011, Van Leer Institute, Jerusalem.

Fox, J., & Shulman, C. (2010). Superintelligence does not imply benevolence. In K. Mainzer (Ed.), *Proceedings of the VIII European Conference on Computing and Philosophy*. Munich: Verlag Dr. Hut.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Essentials of general intelligence: The direct path to artificial general intelligence*. Berlin: Springer.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers, 6*, 31–88.

Gordon-Spears, D. F. (2003). Asimov's Laws: Current Progress. *Lecture Notes in Computer Science, 2699*, 257-259.

Grau, C. (2006). There is no "I" in "Robot": Robots and utilitarianism. *IEEE Intelligent Systems, 21*(4), 52-55.

Guo, S., & Zhang, G. (2009). Robot rights. *Science, 323*(5916), 876.

Hall, J. S. (2007a). *Beyond AI: Creating the conscience of the machine*. Amherst, NY: Prometheus.

Hall, J. S. (2007b). Self-improving AI: An analysis. *Minds and Machines, 17*(3), 249 - 259.

Hanson, R. (2010, Oct. 10). Prefer law to values. *Overcoming Bias* Retrieved Jan. 15, 2012, from http://www.overcomingbias.com/2009/10/prefer-law-to-values.html

Hobbes, T. (1998/1651). *Leviathan*. Oxford: Oxford University Press.

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin: Springer.

Joy, B. (2000, April). Why the future doesn't need us. *Wired Magazine, 8*(4).

Kaczynski, T. (1995, Sep. 19). Industrial society and its future, *The New York Times*.

Kurzweil, R. (2006). *The singularity is near: When humans transcend biology*. New York: Penguin.

LaChat, M. R. (1986). Artificial Intelligence and Ethics: An Exercise in the Moral Imagination. *AI Magazine, 7(2)*, 70-79.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines, 17(4)*, 391-444.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence, 175*(5-6).

McCauley, L. (2007). AI Armageddon and the Three Laws of Robotics. *Ethics and Information Technology, 9(2)*.

McDermott, D. (2008). *Why ethics is a high hurdle for AI*. Paper presented at the North American Conference on Computers and Philosophy, Bloomington, IN.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18-21.

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel & S. Franklin (Eds.), *The proceedings of the first AGI conference* (pp. 483–492). Amsterdam: IOS Press.

Pierce, M. A., & Henry, J. W. (1996). Computer Ethics: The Role of Personal, Informal, and Formal Codes. *Journal of Business Ethics, 14*(4), 425-437.

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems, 21*(4), 46-51.

Pynadath, D. V., & Tambe, M. (2001). *Revisiting Asimov's First Law: A Response to the Call to Arms*. Paper presented at the Intelligent Agents VIII. International Workshop on Agents, Theories, Architectures and Languages (ATAL'01)

Rappaport, Z. H. (2006). Robotics and artificial intelligence: Jewish ethical perspectives. *Acta Neurochirurgica Supplementum, 98*, 9-12.

Ruvinsky, A. I. (2007). Computational ethics. In M. Quigley (Ed.), *Encyclopedia of information ethics and security* (pp. 76-73). Hershey, PA: IGI Global.

Salamon, A., Rayhawk, S., & Kramár, J. (2010). How intelligible is intelligence? In K. Mainzer (Ed.), *Proceedings of the VIII European Conference on Computing and Philosophy*. Munich: Verlag Dr. Hut.

Sawyer, R. J. (2007). Robot ethics. *Science, 318*(5853), 1037.

Sharkey, N. (2008). The ethical frontiers of robotics. *Science, 322*(5909), 1800-1801.

Sotala, K. (2010). From mostly harmless to civilization-threatening: pathways to dangerous artificial general intelligences. In K. Mainzer (Ed.), *Proceedings of the VIII European Conference on Computing and Philosophy*. Munich: Verlag Dr. Hut.

Sotala, K. (2012). Relative advantages of uploads, artificial general intelligences, and other digital minds. *International Journal of Machine Consciousness, 4*.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62-77.

Tonkens, R. (2009). A challenge for machine ethics. *Minds & Machines, 19*(3), 421-438.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, J. Tooby & L. Cosmides (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19-136). Oxford, UK: Oxford University Press.

Veruggio, G. (2010). Roboethics. *IEEE Robotics & Automation Magazine, 17*(2), 105-109.

Wallach, W., & Allen, C. (2006). *EthicALife: A new field of inquiry*. Paper presented at the AnAlifeX workshop, USA.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford, UK: Oxford University Press.

Warwick, K. (2003). Cyborg Morals, Cyborg Values, Cyborg Ethics. *Ethics and Information Technology, 5*, 131-137.

Weld, D. S., & Etzioni, O. (1994). *The First Law of Robotics (a Call to Arms)*. Paper presented at the Twelfth National Conference on Artificial Intelligence (AAAI).

Wright, R. (2001). *Nonzero: The logic of human destiny*. New York: Vintage.

Yampolskiy, R. V. (2011, Oct. 3-4). *Artificial intelligence safety engineering: Why machine ethics is a wrong approach.* Paper presented at the Philosophy and Theory of Artifical Intelligence, Thessaloniki, Greece.

Yampolskiy, R. V. (2013). Turing Test as a Defining Feature of AI-Completeness *Artificial Intelligence, Evolutionary Computation and Metaheuristics - In the footsteps of Alan Turing. Xin-She Yang (Ed.)* (pp. 3-17): Springer.

Yampolskiy, R. V. (October 3-4, 2011). *What to Do with the Singularity Paradox?* Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011), Thessaloniki, Greece.

Yampolskiy, R. V., & Fox, J. (2012a). Artificial Intelligence and the Human Mental Model. In A. Eden, J. Moor, J. Soraker & E. Steinhart (Eds.), *In the Singularity Hypothesis: a Scientific and Philosophical Assessment*: Springer.

Yampolskiy, R. V., & Fox, J. (2012b). Safety Engineering for Artificial General Intelligence. *Topoi. Special Issue on Machine Ethics & the Ethics of Building Intelligent Machines*.

Yudkowsky, E. (2007). The logical fallacy of generalization from fictional evidence. *Less Wrong* Retrieved Feb. 20., 2012, from http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308-345). Oxford: Oxford University Press.

Yudkowsky, E., & Bostrom, N. (2011). The ethics of artificial intelligence. In W. Ramsey & K. Frankish (Eds.), *Cambridge Handbook of Artificial Intelligence*. Cambridge, UK: Cambridge University Press.