The Singularity Institute

Safety Engineering for Artificial General Intelligence

Roman V. Yampolskiy
University of Louisville

Joshua Fox
Singularity Institute Research Associate

Abstract

Machine ethics and robot rights are quickly becoming hot topics in artificial intelligence and robotics communities. We will argue that attempts to attribute moral agency and assign rights to all intelligent machines are misguided, whether applied to infrahuman or superhuman AIs, as are proposals to limit the negative effects of AIs by constraining their behavior. As an alternative, we propose a new science of safety engineering for intelligent artificial agents based on maximizing for what humans value. In particular, we challenge the scientific community to develop intelligent systems that have human-friendly values that they provably retain, even under recursive self-improvement.

Yampolskiy, Roman V., and Joshua Fox. Forthcoming.
Safety engineering for artificial general intelligence. *Topoi*.

This version contains minor changes.

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, and Smit 2006; Moor 2006; Anderson and Anderson 2007; Hall 2007a; McDermott 2008; Tonkens 2009), computer ethics (Pierce and Henry 1996), robot ethics (Sawyer 2007; Sharkey 2008; Lin, Abney, and Bekey 2011), ethicALife (Wallach and Allen 2006), machine morals (Wallach and Allen 2009), cyborg ethics (Warwick 2003), computational ethics (Ruvinsky 2007), roboethics (Veruggio 2010), robot rights (Guo and Zhang 2009), artificial morals (Allen, Smit, and 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 (Asimov 1942; Anderson and Anderson 2007), 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, forthcoming). 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."[1]

[1] 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 (Voss 2007). In this article, unless specified otherwise, we use AI and AGI to refer to artificial general intelligences in the broader sense.

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 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 2005). 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, forthcoming).

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 and 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 species developed (Tooby and Cosmides 1992; Wright 2001; Churchland 2011). Most possible mind architectures and goal systems are profoundly non-anthropomorphic (where "anthropomorphic," 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 and Fox, forthcoming; Muehlhauser and Helm, forthcoming). 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 (LaChat 1986; Weld and Etzioni 1994; Pynadath and Tambe 2002; Gordon-Spears 2003; McCauley 2007). The original laws as given by Asimov (1942) are:

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 (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 humanity (Yampolskiy 2011b).

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; Bostrom and Yudkowsky, forthcoming).

Even if we are successful at designing machines capable of passing a Moral Turing Test (Allen, Varner, and 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, Varner, and Zinser 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 Hanson (2009):

> 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 and 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 [1651] 1998), 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, and Kramár 2010; Sotala, forthcoming). 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. (Bostrom and Yudkowsky, forthcoming)

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.

Moreover, given the complexity of human values, specifying a single desirable value is insufficient to guarantee an outcome positive for humans. Outcomes in which a single value is highly optimized while other values are neglected tend to be disastrous for humanity, as for example one in which a happiness-maximizer turns humans into passive recipients of an electrical feed into pleasure centers of the brain. For a positive outcome, it is necessary to define a goal system that takes into account the entire ensemble of human values simultaneously (Yudkowsky 2011a).

In summary, the ethical principles of give and take, of human motivations constrained by the needs of other humans, and of morality as a necessarily in-built terminal value, need not apply to a non-anthropomorphic superintelligence with arbitrary goals. Safety engineering is needed.

3. AI Safety Engineering

We propose that philosophical discussions of ethics for machines be expanded from today's infrahuman AIs to include artificial general intelligences, and in particular superhuman intelligences. On the theoretical plane, this is important because of the philosophical implications of non-anthropomorphic agents. On the practical plane, given that such AIs may be created within decades (Bostrom 2006), it is essential to supplement philosophy with applied science and engineering aimed at creating safe machines: a new field which we will term "AI Safety Engineering." For brain-inspired AIs, the focus will be on preserving the essential humanity of their values, without allowing moral corruption or technical hardware and software corruption to change them for the worse. For de novo AIs, the focus will be in defining goal systems that help humanity, and then preserving those goals under recursive self-improvement toward superintelligence.

Some work in this important area has already begun (Gordon 1998; Gordon-Spears 2003; Spears 2006). A common theme in AI safety research is the possibility of keeping a superintelligent agent in sealed hardware in order to prevent it from doing harm to humankind. Drexler (1986) suggested confining transhuman machines so that their outputs could be studied and used safely. Chalmers (2010) described the idea of a "leakproof singularity" ("singularity" in the sense of "AI at human level and above"). He suggests that for safety reasons, AIs first be restricted to simulated virtual worlds until their behavioral tendencies can be fully understood under the controlled conditions. Armstrong, Sandberg, and Bostrom (forthcoming) expand on this concept, referred to as "AI Boxing," and further propose an idea for an Oracle AI, which would be only capable of answering questions, rather than taking practical action.

Such confinement is so challenging as to be considered by some impossible. A greater-than-human intelligence would be able to outwit any human gatekeeper, convincing him to take actions that will "release" it (Yudkowsky 2002).

Further substance is added to the concept of AI Boxing with a proposal for a formalized AI confinement protocol representing the confinement of the AI as a computer security challenge (Yampolskiy 2012b). The Artificial Intelligence Confinement Problem is defined 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, unless such information exchange is authorized by the confinement authority. An AI system that succeeds in violating the Confinement Problem protocol is said to have escaped.

One variant of the proposed protocol is based on the idea of asking "safe questions" with restricted answers. A safe question, for this purpose, is one for which predefined options exist, and for which a human being could verify an answer without any help from the superintelligence. Why is this useful? Suppose that a scientist working on a cure for cancer has considered two potential cures. Each one could be developed and tested without assistance from the AI, but would take about three years of hard work each. If he believes that they have about an equal chance of working, the scientist does not know which one to try first. He could ask the AI to suggest which option to try first. Even if a wrong option is chosen to be tested first, humanity will still obtain a cure for cancer, but it will happen three years later. This question is safe because a human being developed the solution, and so the solution cannot be a ruse by the AI to escape. Asking the AI for help will only speed up the progress, but in no way change the outcome. The answer to this question could be encoded in a single bit per proposed option, making it hard to hide additional messages in the provided answer (Yampolskiy 2012b).

To make sure that a potential question is safe, a group of experts could review it. The experts should 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 as applied to generally capable AIs. Such a review will not guarantee safety—if an AI surpasses human levels of intelligence, it can ipso facto outwit the humans—but will reduce the risk when the AI is at infrahuman levels and perhaps minimize the dangers of an ascent through self-improvement.

Even this AI confinement protocol is not perfectly safe, since complete physical isolation is impossible. For example, the miniscule electrical currents needed for the AIs implementation constitute an interaction with the outside world (Vassar 2005; Yampolskiy 2012b). Such currents may seem incapable of significant effect from our perspective, but an intelligence smarter than us may have other ideas.

Any AI Boxing strategy might fail against a superintelligence and must be considered a stopgap measure rather than a guaranteed protection of safety. Nonetheless, in the interim, until superhuman intelligence arises, such a strategy may prove useful.

An existing discipline that might be valuable for training of AI review experts is artimetrics (Yampolskiy and Govindaraju 2008; Gavrilova and Yampolskiy 2011; Yampolskiy and Gavrilova, forthcoming) which identifies, classifies and authenticates AI agents, robots, and virtual reality avatars for security purposes.[2] Extending technologies such as CAPTCHAs, which attempt to distinguish human from robotic visitors to a website (von Ahn et al. 2003; Yampolskiy 2012a), artimetrics takes an adversarial approach to this problem, overcoming attempts to disguise the identities of software agents.

[2] The term “artimetrics” was coined (Yampolskiy and Govindaraju 2008) on the basis of “artilect,” which is Hugo de Garis’s (2005) neologism for “artificial intellect.”

Malware includes some of the most powerful AI technology known today. By applying artimetric techniques to narrow AIs, and gradually building out the techniques in response to improvements by adversaries, artimetrics may evolve into a methodology capable of contending with yet more powerful AIs.

4. Grand Challenge

As the grand challenge of AI safety engineering, we propose the problem of developing safety mechanisms for self-improving systems. 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. Such a mechanism cannot be a rule or constraint on the behavior of the AI in attempting to achieve its goals, since superintelligent agents can probably outwit every constraint imposed by humans. Rather, the AI must *want* to cooperate—it must have safe and stable end-goals from the beginning. Ideally, every generation of a self-improving system should be able to produce a verifiable proof of its safety and the safety of any upgrade 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 not solvable, or that if it is solvable, that it will not be possible to prove that the discovered solution is correct (de Garis 2005; Legg 2006; Goertzel 2011). Extrapolating from the human example has limitations, but it appears that for practical intelligence, overcoming combinatorial explosions in problem solving can only be done by creating complex subsystems optimized for specific challenges. As the complexity of any system increases, the number of errors in the design increases proportionately or perhaps even exponentially, rendering self-verification impossible. Self-improvement radically increases the difficulty, since self-improvement requires reflection, and today's decision theories fail many reflective problems. A single bug in such a system would negate any safety guarantee. Given the tremendous implications of failure, the system must avoid not only bugs in its construction, but also bugs introduced even after the design is complete, whether 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.

The mathematical difficulties of formalizing such safety are imposing. Löb's Theorem, which states that a consistent formal system cannot prove in general that it is sound, may make it impossible for an AI to prove safety properties about itself or a potential new generation of AI (Yudkowsky 2011b). Contemporary decision theories fail on recursion, i.e., in making decisions that depend on the state of the decision system itself. Though tentative efforts are underway to resolve this (Drescher 2006; Yudkowsky 2010), the state of the art leaves us unable to prove goal preservation formally. On the other hand, there will be a powerful agent helping to preserve the AI's goals: the AI itself (Omohundro 2008).

5. 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 et al. 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 superintelligent 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 machine's decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better 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 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 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."

6. Robot Rights

Lastly, we would like to address a sub-branch of machine ethics that, on the surface, has little to do with safety, but that is raised in connection to decisions about future intelligent machines: robot rights (Roth 2009). The question is whether our mind children should automatically be given rights, privileges and responsibilities enjoyed by those granted personhood by society. We believe that, unless such mind children have those characteristics that give humans status as moral agents and/or patients, the answer is "no." While the consensus that all humans are equal in moral status benefits human society and individuals, intelligent machines designed to serve us should not be designed to have human-like characteristics. They should not desire freedom, social status, and other human values; they should not feel suffering and pain as qualia (Dennett 1978; Bishop 2009); and in general they should not have those features that make us ascribe rights to humans. Such intelligent machines should be built entirely to serve human goals; indeed, this is almost a tautology. One might ask of those who think that intelligent machines should always have the features that entail deserving rights: What human goals are served by avoiding making non-person intelligent machines that would otherwise benefit humans? In short, intelligent machines should be built as tools, albeit tools with optimization power, "intelligence," much stronger than ours.

To go one step further, it might be best not to make AIs extremely human-like in appearance, to avoid erroneous attributions that may blur the bright lines we set around moral categories (Arneson 1999). If such confusion were to develop, given the strong human tendency to anthropomorphize, we might encounter rising social pressure to give robots civil and political rights, as an extrapolation of the universal consistency that has proven so central to ameliorating the human condition. Since artificial minds on a general-purpose computing infrastructure can be duplicated easily, and since conversely they can link up to each other with a degree of cohesion unparalleled in humans, extending human-like rights arbitrarily to machine minds would lead to a breakdown of a political system designed to, among other things, help humans get along with each other.

7. 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 non-minds, but in all cases with a careful consideration of risks and security protocols (Yampolskiy 2011a).

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 (Moor 2006; Anderson and Anderson 2007; Tonkens 2009; Lin, Abney, and Bekey 2011) 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.

On the other hand, ethical issues with AIs at human level and above have been addressed by a handful of philosophers, but mostly in the domain of science fiction. Perhaps because of advocacy by organizations like the Singularity Institute and the Future of Humanity Institute at Oxford University, the topic of safety of AIs at human levels of intelligence and above has slowly started to appear in mainstream AI publications.

We call on authors and readers of this volume to start specialized peer-reviewed journals and conferences devoted to the ethics of future artificial general intelligence, These 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.

As far back as 1863, Samuel Butler, best known for his utopian novel *Erewhon* (Butler 1872) published a foresightful article "Darwin Among the Machines," in which he explore the implications of growing machine capabilities (Butler 1863):

> We refer to the question: What sort of creature man's next successor in the supremacy of the earth is likely to be. We have often heard this debated; but it appears to us that we are ourselves creating our own successors; we are daily adding to the beauty and delicacy of their physical organisation; we are daily giving them greater power and supplying by all sorts of ingenious contrivances that self-regulating, self-acting power which will be to them what intellect has been to the human race. In the course of ages we shall find ourselves the inferior race.

Butler had the first inklings of the challenge ahead of us, as we develop our mind children towards intelligence equal to and superior to our own. He did not imagine, however, the risks posed by an intelligence that improves itself to levels so much beyond ours that we become not just an "inferior race," but destroyed as a side-effect of the entity's activities in pursuit of its goals.

Acknowledgments

This article is an expanded version of the conference paper "Artificial intelligence safety engineering: Why machine ethics is a wrong approach" (Yampolskiy 2011a). We would like to thank Brian Rabkin and Michael Anissimov for their comments.

References

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. In Ethics of new information technology papers from CEPE 2005. *Ethics and Information Technology* 7 (3): 149–155. doi:10.1007/s10676-006-0004-4.

Allen, Colin, Gary Varner, and Jason Zinser. 2000. Prolegomena to any future artificial moral agent. In Philosophical foundations of artificial intelligence. Special issue, *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3): 251–261. doi:10.1080/09528130050111428.

Allen, Colin, Wendell Wallach, and Iva Smit. 2006. Why machine ethics? *IEEE Intelligent Systems* 21 (4): 12–17. doi:10.1109/MIS.2006.83.

Anderson, Michael, and Susan Leigh Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28 (4): 15–26. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2065/2052.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. Forthcoming. Thinking inside the box: Using and controlling an oracle AI. *Minds and Machines.*

Arneson, Richard J. 1999. What, if anything, renders all humans morally equal? In *Singer and his critics,* ed. Dale Jamieson. Philosophers and Their Critics 8. Malden, MA: Blackwell.

Asimov, Isaac. 1942. Runaround. *Astounding Science-Fiction,* Mar., 94–103.

Berg, Paul, David Baltimore, Sydney Brenner, Richard O. Roblin, and Maxine F. Singer. 1975. Summary statement of the Asilomar conference on recombinant DNA molecules. *Proceedings of the National Academy of Sciences* 72 (6): 1981–1984. doi:10.1073/pnas.72.6.1981.

Bishop, Mark. 2009. Why computers can't feel pain. In Computation and the natural world, ed. Colin T. A. Schmidt. Special issue, *Minds and Machines* 19 (4): 507–516. doi:10.1007/s11023-009-9173-3.

Bostrom, Nick. 2002. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology* 9. http://www.jetpress.org/volume9/risks.html.

———. 2006. How long before superintelligence? *Linguistic and Philosophical Investigations* 5 (1): 11–30.

Bostrom, Nick, and Eliezer Yudkowsky. Forthcoming. The ethics of artificial intelligence. In *Cambridge handbook of artificial intelligence,* ed. Keith Frankish and William Ramsey. New York: Cambridge University Press.

Butler, Samuel [Cellarius, pseud.]. 1863. Darwin among the machines. *Christchurch Press,* June 13. http://www.nzetc.org/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html.

———. 1872. *Erewhon; or, Over the range.* London: Trübner.

Chalmers, David John. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

Churchland, Patricia S. 2011. *Braintrust: What neuroscience tells us about morality.* Princeton, NJ: Princeton University Press.

Clarke, Roger. 1993. Asimov's laws of robotics: Implications for information technology, part 1. *Computer* 26 (12): 53–61. doi:10.1109/2.247652.

———. 1994. Asimov's laws of robotics: Implications for information technology, part 2. *Computer* 27 (1): 57–66. doi:10.1109/2.248881.

de Garis, Hugo. 2005. *The artilect war: Cosmists vs. terrans: A bitter controversy concerning whether humanity should build godlike massively intelligent machines.* Palm Springs, CA: ETC Publications.

Dennett, Daniel C. 1978. Why you can't make a computer that feels pain. In Automaton-theoretical foundations of psychology and biology, part I. *Synthese* 38 (3): 415–456. doi:10.1007/BF00486638.

Drescher, Gary L. 2006. *Good and real: Demystifying paradoxes from physics to ethics.* Bradford Books. Cambridge, MA: MIT Press.

Drexler, K. Eric. 1986. *Engines of creation.* Garden City, NY: Anchor Press.

Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. Forthcoming. *The singularity hypothesis: A scientific and philosophical assessment.* Berlin: Springer.

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

Fox, Joshua, and Carl Shulman. 2010. Superintelligence does not imply benevolence. In Mainzer 2010.

Gauthier, David P. 1986. *Morals by agreement.* New York: Oxford University Press. doi:10.1093/0198249926.001.0001.

Gavrilova, Marina L., and Roman V. Yampolskiy. 2011. Applying biometric principles to avatar recognition. In *Transactions on computational science XII: Special issue on cyberworlds,* ed. Marina L. Gavrilova, C. J. Kenneth Tan, Alexei Sourin, and Olga Sourina, 140–158. Lecture Notes in Computer Science 6670. Berlin: Springer. doi:10.1007/978-3-642-22336-5_8.

Goertzel, Ben. 2011. Does humanity need an AI nanny? *H+ Magazine,* Aug. 17. http://hplusmagazine.com/2011/08/17/does-humanity-need-an-ai-nanny/.

Good, Irving John. 1965. Speculations concerning the first ultraintelligent machine. In *Advances in computers,* ed. Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Gordon, Diana F. 1998. Well-behaved borgs, bolos, and berserkers. In *Proceedings of the 15th international conference on machine learning (ICML-98),* ed. Jude W. Shavlik, 224–232. San Francisco, CA: Morgan Kaufmann.

Gordon-Spears, Diana F. 2003. Asimov's laws: Current progress. In *Formal approaches to agent-based systems: Second international workshop, FAABS 2002, Greenbelt, MD, USA, October 29–31, 2002. Revised papers,* ed. Michael G. Hinchey, James L. Rash, Walter F. Truszkowski, Christopher Rouff, and Diana F. Gordon-Spears, 257–259. Lecture Notes in Computer Science 2699. Berlin: Springer. doi:10.1007/978-3-540-45133-4_23.

Grau, Christopher. 2006. There is no "I" in "Robot": Robots and utilitarianism. *IEEE Intelligent Systems* 21 (4): 52–55. doi:10.1109/MIS.2006.81.

Guo, Shesen, and Ganzhou Zhang. 2009. Robot rights. *Science* 323 (5916): 876. doi:10.1126/science.323.5916.876a.

Hall, John Storrs. 2007a. *Beyond AI: Creating the conscience of the machine.* Amherst, NY: Prometheus Books.

———. 2007b. Self-improving AI: An analysis. *Minds and Machines* 17 (3): 249–259. doi:10.1007/s11023-007-9065-3.

Hanson, Robin. 2009. Prefer law to values. Overcoming Bias (blog). Oct. 10. http://www.overcomingbias.com/2009/10/prefer-law-to-values.html (accessed Mar. 26, 2012).

Hobbes, Thomas. [1651] 1998. *Leviathan.* Oxford World's Classics. Repr. New York: Oxford University Press.

Hutter, Marcus. 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability.* Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

Joy, Bill. 2000. Why the future doesn't need us. *Wired,* Apr. http://www.wired.com/wired/archive/8.04/joy.html.

Kaczynski, Theodore. 1995. Industrial society and its future. *Washington Post,* Sept. 19.

Kurzweil, Ray. 2005. *The singularity is near: When humans transcend biology.* New York: Viking.

LaChat, Michael R. 1986. Artificial intelligence and ethics: An exercise in the moral imagination. *AI Magazine* 7 (2): 70–79. http://www.aaai.org/ojs/index.php/aimagazine/article/view/540/476.

Legg, Shane. 2006. Unprovability of friendly AI. Vetta Project (blog). Sept. 15. http://www.vetta.org/2006/09/unprovability-of-friendly-ai/ (accessed Jan. 15, 2012).

Legg, Shane, and Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17 (4): 391–444. doi:10.1007/s11023-007-9079-x.

Lin, Patrick, Keith Abney, and George Bekey. 2011. Robot ethics: Mapping the issues for a mechanized world. Ed. Randy Goebel and Mary-Anne Williams. Special review issue, *Artificial Intelligence* 175 (5–6): 942–949. doi:10.1016/j.artint.2010.11.026.

Mainzer, Klaus, ed. 2010. *ECAP10: VIII European Conference on Computing and Philosophy.* Munich: Verlag Dr. Hut.

McCauley, Lee. 2007. AI armageddon and the three laws of robotics. *Ethics and Information Technology* 9 (2): 153–164. doi:10.1007/s10676-007-9138-2.

McDermott, Drew. 2008. Why ethics is a high hurdle for AI. Paper presented at the 2008 North American Conference on Computing and Philosophy, Indiana University, Bloomington, July 10–12. http://cs-www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf (accessed May 18, 2012).

Moor, James H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21 (4): 18–21. doi:10.1109/MIS.2006.80.

Muehlhauser, Luke, and Louie Helm. Forthcoming. The singularity and machine ethics. In Eden, Søraker, Moor, and Steinhart, forthcoming.

Omohundro, Stephen M. 2008. The basic AI drives. In *Artificial general intelligence 2008: Proceedings of the first AGI conference,* ed. Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS Press.

Pierce, Margaret Anne, and John W. Henry. 1996. Computer ethics: The role of personal, informal, and formal codes. *Journal of Business Ethics* 15 (4): 425–437. doi:10.1007/BF00380363.

Powers, Thomas M. 2006. Prospects for a Kantian machine. *IEEE Intelligent Systems* 21 (4): 46–51. doi:10.1109/MIS.2006.77.

Pynadath, David V., and Milind Tambe. 2002. Revisiting Asimov's first law: A response to the call to arms. In *Intelligent agents VIII: Agent theories, architectures, and languages 8th international workshop, ATAL 2001 Seattle, WA, USA, August 1—3, 2001 Revised Papers,* ed. John-Jules Ch. Meyer and Milind Tambe, 307–320. Berlin: Springer. doi:10.1007/3-540-45448-9_22.

Rappaport, Z. H. 2006. Robotics and artificial intelligence: Jewish ethical perspectives. In *Medical technologies in neurosurgery,* ed. Christopher Nimsky and Rudolf Fahlbusch, 9–12. Acta Neurochirurgica Supplementum 98. Vienna: Springer. doi:10.1007/978-3-211-33303-7_2.

Roth, Daniel. 2009. Do humanlike machines deserve human rights? *Wired,* Jan. 19. http://www.wired.com/culture/culturereviews/magazine/17-02/st_essay.

Ruvinsky, Alicia I. 2007. Computational ethics. In *Encyclopedia of information ethics and security,* ed. Marian Quigley, 76–82. IGI Global. doi:10.4018/978-1-59140-987-8.ch012.

Salamon, Anna, Stephen Rayhawk, and János Kramár. 2010. How intelligible is intelligence? In Mainzer 2010.

Sawyer, Robert J. 2007. Robot ethics. *Science* 318 (5853): 1037. doi:10.1126/science.1151606.

Sharkey, Noel. 2008. The ethical frontiers of robotics. *Science* 322 (5909): 1800–1801. doi:10.1126/science.1164582.

Sotala, Kaj. 2010. From mostly harmless to civilization-threatening: Pathways to dangerous artificial intelligences. In Mainzer 2010.

———. Forthcoming. Advantages of artificial intelligences, uploads, and digital minds. *International Journal of Machine Consciousness* 4. Preprint at, http://www.xuenay.net/Papers/DigitalAdvantages.pdf.

Sparrow, Robert. 2007. Killer robots. *Journal of Applied Philosophy* 24 (1): 62–77. doi:10.1111/j.1468-5930.2007.00346.x.

Spears, Diana F. 2006. Assuring the behavior of adaptive agents. In *Agent technology from a formal perspective,* ed. Christopher Rouff, Michael Hinchey, James Rash, Walter Truszkowski, and Diana F. Gordon-Spears, 227–257. NASA Monographs in Systems and Software Engineering. London: Springer. doi:10.1007/1-84628-271-3_8.

Tonkens, Ryan. 2009. A challenge for machine ethics. *Minds and Machines* 19 (3): 421–438. doi:10.1007/s11023-009-9159-1.

Tooby, John, and Leda Cosmides. 1992. The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture,* ed. Jerome H. Barkow, Leda Cosmides, and John Tooby, 19–136. New York: Oxford University Press.

Vassar, Michael. 2005. AI boxing (dogs and helicopters). SL4. Aug. 2. http://sl4.org/archive/0508/11817.html (accessed Jan. 18, 2012).

Veruggio, Gianmarco. 2010. Roboethics. *IEEE Robotics & Automation Magazine,* June, 105–109. doi:10.1109/MRA.2010.936959.

von Ahn, Luis, Manuel Blum, Nicholas J. Hopper, and John Langford. 2003. CAPTCHA: Using hard AI problems for security. In *Advances in cryptology — EUROCRYPT 2003: International conference on the theory and applications of cryptographic techniques, Warsaw, Poland, May 4-8, 2003 proceedings,* ed. Eli Biham, 293–311. Lecture Notes in Computer Science 2656. Berlin: Springer. doi:10.1007/3-540-39200-9_18.

Voss, Peter. 2007. Essentials of general intelligence: The direct path to artificial general intelligence. In *Artificial general intelligence,* ed. Ben Goertzel and Cassio Pennachin, 131–157. Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4_4.

Wallach, Wendell, and Colin Allen. 2006. EthicALife: A new field of inquiry. Paper presented at EthicALife: An ALifeX Workshop, Bloomington, IN, June 3–7. http://ethicalife.dynalias.org/Allen-Wallach.pdf.

———. 2009. *Moral machines: Teaching robots right from wrong.* New York: Oxford University Press. doi:10.1093/acprof:oso/9780195374049.001.0001.

Warwick, Kevin. 2003. Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology* 5 (3): 131–137. doi:10.1023/B:ETIN.0000006870.65865.cf.

Weld, Daniel, and Oren Etzioni. 1994. The first law of robotics (a call to arms). In *Proceedings of the twelfth national conference on artificial intelligence,* ed. Barbara Hayes-Roth and Richard E. Korf, 1042–1047. Menlo Park, CA: AAAI Press. http://www.aaai.org/Papers/AAAI/1994/AAAI94-160.pdf.

Wright, Robert. 2001. *Nonzero: The logic of human destiny.* New York: Vintage.

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

———. 2011b. What to do with the singularity paradox? Paper presented at the Philosophy and Theory of Artificial Intelligence (PT-AI 2011), Thessaloniki, Greece, Oct. 3–4.

———. 2012a. AI-complete CAPTCHAs as zero knowledge proofs of access to an artificially intelligent system. *ISRN Artificial Intelligence* 2012:271878. doi:10.5402/2012/271878.

———. 2012b. Leakproofing the singularity: artificial intelligence confinement problem. *Journal of Consciousness Studies* 2012 (1–2): 194–214. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00014.

———. Forthcoming. Turing test as a defining feature of AI-completeness. In *Artificial intelligence, evolutionary computing and metaheuristics: In the footsteps of Alan Turing,* ed. Xin-She Yang. Studies in Computational Intelligence 427. Berlin: Springer.

Yampolskiy, Roman V., and Joshua Fox. Forthcoming. Artificial general intelligence and the human mental model. In Eden, Søraker, Moor, and Steinhart, forthcoming.

Yampolskiy, Roman V., and Marina L. Gavrilova. Forthcoming. Artimetrics: Biometrics for artificial entities. *IEEE Robotics & Automation Magazine.*

Yampolskiy, Roman V., and Venu Govindaraju. 2008. Behavioral biometrics for verification and recognition of malicious software agents. In *Sensors, and command, control, communications, and intelligence (C3I) technologies for homeland security and homeland defense VII: 17–20 March 2008, Orlando, Florida, USA,* ed. Edward M. Carapezza, 694303. Proceedings of SPIE 6943. Bellingham, WA: SPIE. doi:10.1117/12.773554.

Yudkowsky, Eliezer. 2002. The AI-box experiment. http://yudkowsky.net/singularity/aibox (accessed Jan. 15, 2012).

———. 2007. The logical fallacy of generalization from fictional evidence. LessWrong. Oct. 16. http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/ (accessed Feb. 20, 2012).

———. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks,* ed. Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

———. 2010. *Timeless decision theory.* The Singularity Institute, San Francisco, CA. http://singinst.org/upload/TDT-v01o.pdf.

———. 2011a. Complex value systems in friendly AI. In *Artificial general intelligence: 4th international conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings,* ed. Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388–393. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_48.

———. 2011b. Open problems in friendly artificial intelligence. Paper presented at Singularity Summit 2011, New York, Oct. 15–16. http://www.youtube.com/watch?v=MwriJqBZyoM.