
# Responses to Catastrophic AGI Risk: A Survey

Kaj Sotala
*Machine Intelligence Research Institute*

Roman V. Yampolskiy
*University of Louisville*

## Abstract

Many researchers have argued that humanity will create artificial general intelligence (AGI) within the next twenty to one hundred years. It has been suggested that AGI may pose a catastrophic risk to humanity. After summarizing the arguments for why AGI may pose such a risk, we survey the field's proposed responses to AGI risk. We consider societal proposals, proposals for external constraints on AGI behaviors, and proposals for creating AGIs that are safe due to their internal design.

Sotala, Kaj and Roman V. Yampolskiy. 2013. *Responses to Catastrophic AGI Risk: A Survey*.
Technical report 2013-2. Berkeley, CA: Machine Intelligence Research Institute.
Last modified September 13, 2013.

# Contents

## 1. Introduction[1]

Many philosophers, futurologists, and artificial intelligence researchers have argued that in the next twenty to one hundred years we will create artificial general intelligences, or AGIs (Moravec 1988, 1999; Bostrom 1998; Warwick 1998; Kurzweil 2005; Baum, Goertzel, and Goertzel 2011; Sandberg and Bostrom 2011; Muehlhauser and Salamon 2012).[2] Unlike current "narrow" AI systems, AGIs would perform at or above the human level not merely in particular domains (e.g., chess or arithmetic), but in a wide variety of domains, including novel ones.[3] They would have a robust understanding of natural language and be capable of general problem solving.

The creation of AGI could pose challenges and risks of varied severity for society, such as the possibility of AGIs outcompeting humans in the job market (Brynjolfsson and McAfee 2011; Miller 2012). This article, however, focuses on the suggestion that AGIs may come to act in ways not intended by their creators, and in this way pose a *catastrophic* (Bostrom and Ćirković 2008b) or even an *existential* (Bostrom 2002) risk to humanity.[4] We will organize and summarize the proposals that have been made so far for responding to catastrophic AGI risk, so as to provide a map of the field to newcomers and veterans alike.

Section 2 explains why AGI may pose a catastrophic risk. Sections 3–5 survey three categories of proposals for dealing with AGI risk: societal proposals, proposals for external constraints on AGI behaviors, and proposals for creating AGIs that are safe due to their internal design. Although our paper is not intended to be particularly

1. This paper is a greatly expanded and elaborated version of Yampolskiy (2013).

2. For a preliminary "AGI roadmap," see Adams et al. (2012). For a variety of views, see Eden et al. (2012).

The term "AGI" was introduced by Gubrud (1997). For overviews of AGI approaches, see Wang, Goertzel, and Franklin (2008) and Adams et al. (2012). Some closely related terms are "strong AI" (e.g., Kurzweil 2005) and "human-level AI" (e.g., Minsky, Singh, and Sloman 2004; Cassimatis, Mueller, and Winston 2006). Unlike the term "human-level AI," the term "Artificial General Intelligence" does not necessarily presume that the intelligence will be human-like.

3. For this paper, we use a binary distinction between narrow AI and AGI. This is merely for the sake of simplicity: we do not assume the actual difference between the two categories to necessarily be so clean-cut.

4. A catastrophic risk is something that might inflict serious damage to human well-being on a global scale and cause ten million or more fatalities (Bostrom and Ćirković 2008b). An existential risk is one that threatens human extinction (Bostrom 2002). Many writers argue that AGI might be a risk of such magnitude (Butler 1863; Wiener 1960; Good 1965; Cade 1966; Moravec 1988; Vinge 1993; Warwick 1998; Joy 2000; Bostrom 2002; de Garis 2005; Hall 2007b; Yudkowsky 2008a; Berglas 2012; Chalmers 2010; Miller 2012; Guterl 2012; Casti 2012).

argumentative, we briefly give our own opinion in each major subsection of sections 3–5 and recap our favorite proposals in section 6.

Attempting to influence the course of developing technologies involves a great deal of uncertainty concerning the nature of those technologies and the likely long-term impacts of societal decisions (Bostrom 2007). While we aim to provide a preliminary analysis of many proposals, we do not have final answers. The purpose of this paper is to act as a survey and to highlight some considerations that will be useful starting points for further research.

In the hopes of fostering further debate, we also highlight some of the proposals that we consider the most promising. In the medium term, these are regulation (section 3.3), merging with machines (section 3.4), AGI confinement (section 4.1), Oracle AI (section 5.1), and motivational weaknesses (section 5.6). In the long term, the most promising approaches seem to be value learning (section 5.2.5) and human-like architectures (section 5.3.4). Section 6 provides an extended discussion of the various merits and problems of these proposals.

## 2. Catastrophic AGI Risk

We begin with a brief sketch of the argument that AGI poses a catastrophic risk to humanity. At least two separate lines of argument seem to support this conclusion.

First, AI has already made it possible to automate many jobs (Brynjolfsson and McAfee 2011), and AGIs, when they are created, should be capable of performing *most* jobs better than humans (Warwick 1998; Hanson 2008; Hall 2008; Miller 2012). As humanity grows increasingly reliant on AGIs, these AGIs will begin to wield more and more influence and power. Even if AGIs initially function as subservient tools, an increasing number of decisions will be made by autonomous AGIs rather than by humans. Over time it would become ever more difficult to replace the AGIs, even if they no longer remained subservient.

Second, there may be a sudden discontinuity in which AGIs rapidly become far more numerous or intelligent (Good 1965; Solomonoff 1985; Yudkowsky 1996; Shulman and Sandberg 2010; Chalmers 2010). This could happen due to (1) a conceptual breakthrough which makes it easier to run AGIs using far less hardware, (2) AGIs using fast computing hardware to develop ever-faster hardware, or (3) AGIs crossing a threshold in intelligence that allows them to carry out increasingly fast software self-improvement. Even if the AGIs were expensive to develop at first, they could be cheaply copied and could thus spread quickly once created.

Once they become powerful enough, AGIs might be a threat to humanity even if they are not actively malevolent or hostile. Mere indifference to human values—including

human survival—could be sufficient for AGIs to pose an existential threat (Yudkowsky 2008a, 2011; Omohundro 2007, 2008).

We will now lay out the above reasoning in more detail.

### 2.1. Most Tasks Will Be Automated

Ever since the Industrial Revolution, society has become increasingly automated. Brynjolfsson and McAfee (2011) argue that the current high unemployment rate in the United States is partially due to rapid advances in information technology, which has made it possible to replace human workers with computers faster than human workers can be trained in jobs that computers cannot yet perform. Vending machines are replacing shop attendants, automated discovery programs which locate relevant legal documents are replacing lawyers and legal aides, and automated virtual assistants are replacing customer service representatives.

Labor is becoming automated for reasons of cost, efficiency, and quality. Once a machine becomes capable of performing a task as well as (or almost as well as) a human, the cost of purchasing and maintaining it may be less than the cost of having a salaried human perform the same task. In many cases, machines are also capable of doing the same job faster, for longer periods, and with fewer errors. In addition to replacing workers entirely, machines may also take over aspects of jobs that were once the sole domain of highly trained professionals, making the job easier to perform by less-skilled employees (Whitby 1996).

If workers can be affordably replaced by developing more sophisticated AI, there is a strong economic incentive to do so. This is already happening with narrow AI, which often requires major modifications or even a complete redesign in order to be adapted for new tasks. "A Roadmap for US Robotics" (Hollerbach, Mason, and Christensen 2009) calls for major investments into automation, citing the potential for considerable improvements in the fields of manufacturing, logistics, health care, and services. Similarly, the US Air Force Chief Scientist's (2010) "Technology Horizons" report mentions "increased use of autonomy and autonomous systems" as a key area of research to focus on in the next decade, and also notes that reducing the need for manpower provides the greatest potential for cutting costs. In 2000, the US Congress instructed the armed forces to have one third of their deep strike force aircraft be unmanned by 2010, and one third of their ground combat vehicles be unmanned by 2015.[5]

To the extent that an AGI could learn to do many kinds of tasks—or even *any* kind of task—without needing an extensive re-engineering effort, the AGI could make the replacement of humans by machines much cheaper and more profitable. As more

5. National Defense Authorization, Fiscal Year 2001, Pub. L. No. 106–398, 114 Stat. 1654 (2000).

tasks become automated, the bottlenecks for further automation will require adaptability and flexibility that narrow-AI systems are incapable of. These will then make up an increasing portion of the economy, further strengthening the incentive to develop AGI.

Increasingly sophisticated AI may eventually lead to AGI, possibly within the next several decades (Baum, Goertzel, and Goertzel 2011; Muehlhauser and Salamon 2012). Eventually it will make economic sense to automate all or nearly all jobs (Warwick 1998; Hanson 2008; Hall 2008). As AGIs will possess many advantages over humans (Sotala 2012; Muehlhauser and Salamon 2012), a greater and greater proportion of the workforce will consist of intelligent machines.

### 2.2. AGIs Might Harm Humans

AGIs might bestow overwhelming military, economic, or political power on the groups that control them (Bostrom 2002). For example, automation could lead to an ever-increasing transfer of wealth and power to the owners of the AGIs (Brain 2003; Brynjolfsson and McAfee 2011). AGIs could also be used to develop advanced weapons and plans for military operations or political takeovers (Gubrud 1997; Bostrom 2002; Karnofsky and Tallinn 2011). Some of these scenarios could lead to catastrophic risks, depending on the capabilities of the AGIs and other factors.

Our focus is on the risk from the possibility that AGIs could behave in unexpected and harmful ways, even if the intentions of their owners were benign. Even modern-day narrow-AI systems are becoming autonomous and powerful enough that they sometimes take unanticipated and harmful actions before a human supervisor has a chance to react. To take one example, rapid automated trading was found to have contributed to the 2010 stock market "Flash Crash" (CFTC & SEC 2010).[6] Autonomous systems may also cause people difficulties in more mundane situations, such as when a credit card is automatically flagged as possibly stolen due to an unusual usage pattern (Allen, Wallach, and Smit 2006), or when automatic defense systems malfunction and cause deaths (Shachtman 2007).

As machines become more autonomous, humans will have fewer opportunities to intervene in time and will be forced to rely on machines making good choices. This has prompted the creation of the field of "machine ethics" (Wallach and Allen 2009; Allen, Wallach, and Smit 2006; Anderson and Anderson 2011), concerned with creating AI systems designed to make appropriate moral choices. Compared to narrow-AI systems,

6. On the less serious front, see Eisen (2011) for an amusing example of automated trading going awry.

AGIs will be even more autonomous and capable, and will thus require even more robust solutions for governing their behavior.[7]

If some AGIs were both powerful and indifferent to human values, the consequences could be disastrous. At one extreme, powerful AGIs indifferent to human survival could bring about human extinction. As Yudkowsky (2008a) writes, "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else."

Omohundro (2007, 2008) and Bostrom (2012) argue that standard microeconomic theory prescribes particular instrumental behaviors which are useful for the achievement of almost any set of goals. Furthermore, any agents which do not follow certain axioms of rational behavior will possess vulnerabilities which some other agent may exploit to their own benefit. Thus AGIs which understand these principles and wish to act efficiently will modify themselves so that their behavior more closely resembles rational economic behavior (Omohundro 2012). Extra resources are useful in the pursuit of nearly any set of goals, and self-preservation behaviors will increase the probability that the agent can continue to further its goals. AGI systems which follow rational economic theory will then exhibit tendencies toward behaviors such as self-replicating, breaking into other machines, and acquiring resources without regard for anyone else's safety. They will also attempt to improve themselves in order to more effectively achieve these and other goals, which could lead to rapid improvement even if the designers did not intend the agent to self-improve.

Even AGIs that were explicitly designed to behave ethically might end up acting at cross-purposes to humanity, because it is difficult to precisely capture the complexity of human values in machine goal systems (Yudkowsky 2008a, 2011; Armstrong, Sandberg, and Bostrom 2012; Muehlhauser and Helm 2012).

Muehlhauser and Helm (2012) caution that moral philosophy has found no satisfactory formalization of human values. All moral theories proposed so far would lead to undesirable consequences if implemented by superintelligent machines. For example, a

7. In practice, there have been two separate communities doing research on automated moral decision-making (Muehlhauser and Helm 2012; Allen and Wallach 2012; Shulman, Jonsson, and Tarleton 2009a). The "AGI ethics" community has concentrated specifically on advanced AGIs (e.g., Yudkowsky 2008a; Goertzel 2010b), while the "machine ethics" community typically has concentrated on more immediate applications for current-day AI (e.g., Wallach, Allen, and Smit 2008; Anderson and Anderson 2011). In this paper, we have cited the machine ethics literature only where it seemed relevant, leaving out papers that seemed to be too focused on narrow-AI systems for our purposes. In particular, we have left out most discussions of military machine ethics (Arkin 2009), which focus primarily on the constrained special case of creating systems that are safe for battlefield usage. Note that while the term "machine ethics" is relatively established, "AGI ethics" is not. One proposed alternative name for the "AGI ethics" discipline is "AI safety engineering" (Yampolskiy and Fox 2012; Yampolskiy 2013).

machine programmed to maximize the satisfaction of human (or sentient) preferences would simply modify people's brains to give them desires that are maximally easy to satisfy.

Intuitively, one might say that current moral theories are all *too simple*—even if they seem correct at first glance, they do not actually take into account all the things that we value, and this leads to a catastrophic outcome. This could be referred to as the *complexity of value thesis*. Recent psychological and neuroscientific experiments confirm that human values are highly complex (Muehlhauser and Helm 2012), that the pursuit of pleasure is not the only human value (Kringelbach and Berridge 2009), and that humans are often unaware of their own values (T. D. Wilson 2002; Ferguson, Hassin, and Bargh 2007; Moskowitz, Li, and Kirk 2004).

Still, perhaps powerful AGIs would have desirable consequences so long as they were programmed to respect *most* human values. If so, then our inability to perfectly specify human values in AGI designs need not pose a catastrophic risk. Different cultures and generations have historically had very different values from each other, and it seems likely that over time our values would become considerably different from current-day ones. It could be enough to maintain some small set of core values, though what exactly would constitute a core value is unclear.[8]

Yudkowsky (2011) argues that, due to the fragility of value, the basic problem remains. He argues that, even if an AGI implemented *most* human values, the outcome might still be unacceptable. For example, an AGI which lacked the value of novelty could create a solar system filled with countless minds experiencing one highly optimal and satisfying experience over and over again, never doing or feeling anything else (Yudkowsky 2009).[9]

8. For example, different people may disagree over whether freedom or well-being is a more important value.

9. Miller similarly notes in *Singularity Rising* that, despite a common belief to the contrary, it is impossible to write laws in a manner that would match our stated moral principles without a judge needing to use a large amount of implicit common-sense knowledge to correctly interpret them:

> Let's expand our definition of murder to *deliberately causing someone's death when not acting in self-defense and when not acting to prevent the death of others*. But now you're innocent of murder if you shoot and kill a guard because he was shooting at you to stop you from robbing a bank because in this situation you took another's life to protect your own. To handle this latest hypothetical, we can define murder as *deliberately causing someone's death, unless you acted in self-defense or in the defense of others against some potential harm that you did not initiate*.
>
> But now consider this example: You hate your boss and pray out loud for his death. You genuinely believe that the God of your faith might act on these prayers. Your office mate who heard the prayers tells your boss what you said. The boss comes to your office and fires you. Because he took the time to fire you, he starts his commute home five minutes later than

In this paper, we will frequently refer to the problem of "AGI safety" or "safe AGI," by which we mean the problem of ensuring that AGIs respect human values, or perhaps some extrapolation or idealization of human values.[10] We do not seek to imply that current human values would be the best possible ones, that AGIs could not help us in developing our values further, or that the values of other sentient beings would be irrelevant. Rather, by "human values" we refer to the kinds of basic values that nearly all humans would agree upon, such as that AGIs forcibly reprogramming people's brains, or destroying humanity, would be a bad outcome. In cases where proposals related to AGI risk might change human values in some major but not as obviously catastrophic way, we will mention the possibility of these changes but remain agnostic on whether they are desirable or undesirable.

We conclude this section with one frequently forgotten point: in order to avoid catastrophic risks or worse, it is not enough to ensure that some AGIs are safe. If there are multiple different AGIs operating, some of them safe and some of them not, the unsafe ones could still end up doing considerable amounts of damage. Proposals which seek to solve the issue of catastrophic AGI risk need to also provide some mechanism for ensuring that *most* (or perhaps even "nearly all") AGIs are either created safe or prevented from doing considerable harm.

### 2.3. AGIs May Become Powerful Quickly

There are several reasons why AGIs may quickly come to wield unprecedented power in society. "Wielding power" may mean having direct decision-making power, or it may

---

> he otherwise would. While driving home, your former boss is struck by a bolt of lightning. Had he left work five minutes earlier, he would have been in a different place when the lightning struck and so would not have been killed. Under our last definition, you are guilty because you caused your boss's death and wanted your boss to die. I could, literally, fill up my book with an expansion of this exercise.
>
> Laws shouldn't always be interpreted literally because legislators can't anticipate all possible contingencies. Also, humans' intuitive feel for what constitutes murder goes beyond anything we can commit to paper. The same applies to friendliness." (Miller 2012)

10. Within the AGI ethics literature, safe autonomous AGI is sometimes called "Friendly AI" (Yudkowsky 2001, 2008a, 2011; McGinnis 2010; Goertzel and Pitt 2012; Miller 2012). Yudkowsky (2001) defines "Friendly AI" as "the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals." However, some papers (e.g., Goertzel [2002, 2006]; Waser [2011]) use "Friendly AI" as a narrower term to refer to safe AGI designs as advocated by Yudkowsky (2001, 2008a). These designs have the goal of benefiting humans as their overarching value, from which all the other goals and values of the system are derived. In this paper we use the term "Friendly AI" to refer to Yudkowsky's proposal, and "safe AGI" as our more general term.

mean carrying out human decisions in a way that makes the decision maker reliant on the AGI. For example, in a corporate context an AGI could be acting as the executive of the company, or it could be carrying out countless low-level tasks which the corporation needs to perform as part of its daily operations.

Bugaj and Goertzel (2007) consider three kinds of AGI scenarios: capped intelligence, soft takeoff, and hard takeoff. In a *capped intelligence* scenario, all AGIs are prevented from exceeding a predetermined level of intelligence and remain at a level roughly comparable with humans. In a *soft takeoff* scenario, AGIs become far more powerful than humans, but on a timescale which permits ongoing human interaction during the ascent. Time is not of the essence, and learning proceeds at a relatively human-like pace. In a *hard takeoff* scenario, an AGI will undergo an extraordinarily fast increase in power, taking effective control of the world within a few years or less.[11] In this scenario, there is little time for error correction or a gradual tuning of the AGI's goals.

The viability of many proposed approaches depends on the hardness of a takeoff. The more time there is to react and adapt to developing AGIs, the easier it is to control them. A soft takeoff might allow for an approach of incremental machine ethics (Powers 2011), which would not require us to have a complete philosophical theory of ethics and values, but would rather allow us to solve problems in a gradual manner. However, there are at least three reasons that hard takeoff is also plausible.

The hard takeoff scenarios can be roughly divided into those involving the quantity of hardware (the *hardware overhang* scenario), the quality of hardware (the *speed explosion* scenario), and the quality of software (the *intelligence explosion* scenario). Although we discuss them separately, it seems plausible that several of them could happen simultaneously and feed into each other.

### 2.3.1. Hardware Overhang

Hardware progress may outpace AGI software progress. Contemporary supercomputers already rival or even exceed some estimates of the computational capacity of the human brain, while no software seems to have both the brain's general learning capacity and its scalability.[12] If such trends continue, then by the time the software for AGI is

11. Bugaj and Goertzel defined hard takeoff to refer to a period of months or less. We have chosen a somewhat longer time period, as even a few years might easily turn out to be too little time for society to properly react.

12. Bostrom (1998) estimates that the effective computing capacity of the human brain might be somewhere around $10^{17}$ operations per second (OPS), and Moravec (1998) estimates it at $10^{14}$ OPS. As of November 2012, the fastest supercomputer in the world had achieved a top capacity of $10^{16}$ floating-point operations per second (FLOPS) and the five-hundredth fastest a top capacity of $10^{13}$ FLOPS (Top500

invented there may be a *computing overhang*—an abundance of cheap hardware available for running thousands or millions of AGIs, possibly with a speed of thought much faster than that of humans (Yudkowsky 2008b; Shulman and Sandberg 2010; Sotala 2012).

As increasingly sophisticated AGI software becomes available, it would be possible to rapidly copy improvements to millions of servers, each new version being capable of doing more kinds of work or being run with less hardware. Thus, the AGI software could replace an increasingly large fraction of the workforce.[13] The need for AGI systems to be trained for some jobs would slow the rate of adoption, but powerful computers could allow for fast training. If AGIs end up doing the vast majority of work in society, humans could become dependent on them.

AGIs could also plausibly take control of Internet-connected machines in order to harness their computing power (Sotala 2012); Internet-connected machines are regularly compromised.[14]

### 2.3.2. Speed Explosion

Another possibility is a *speed explosion* (Solomonoff 1985; Yudkowsky 1996; Chalmers 2010; Hutter 2012), in which intelligent machines design increasingly faster machines. A hardware overhang might contribute to a speed explosion, but is not required for it. An AGI running at the pace of a human could develop a second generation of hardware on which it could run at a rate faster than human thought. It would then require a shorter time to develop a third generation of hardware, allowing it to run faster than on the previous generation, and so on. At some point, the process would hit physical limits

2012). Note however that OPS and FLOPS are not directly comparable and there is no reliable way of interconverting the two. Sandberg and Bostrom (2008) estimate that OPS and FLOPS grow at a roughly comparable rate.

13. The speed that would allow AGIs to take over most jobs would depend on the cost of the hardware and the granularity of the software upgrades. A series of upgrades over an extended period, each producing a 1% improvement, would lead to a more gradual transition than a single upgrade that brought the software from the capability level of a chimpanzee to a rough human equivalence. Note also that several companies, including Amazon and Google, offer vast amounts of computing power for rent on an hourly basis. An AGI that acquired money and then invested all of it in renting a large amount of computing resources for a brief period could temporarily achieve a much larger boost than its budget would otherwise suggest.

14. Botnets are networks of computers that have been compromised by outside attackers and are used for illegitimate purposes. Rajab et al. (2007) review several studies which estimate the sizes of the largest botnets as being between a few thousand to 350,000 bots. Modern-day malware could theoretically infect any susceptible Internet-connected machine within tens of seconds of its initial release (Staniford, Paxson, and Weaver 2002). The Slammer worm successfully infected more than 90% of vulnerable hosts within ten minutes, and had infected at least 75,000 machines by the thirty-minute mark (Moore et al. 2003). The previous record holder in speed, the Code Red worm, took fourteen hours to infect more than 359,000 machines (Moore, Shannon, and Brown 2002).

and stop, but by that time AGIs might come to accomplish most tasks at far faster rates than humans, thereby achieving dominance.

The extent to which the AGI needs humans in order to produce better hardware will limit the pace of the speed explosion, so a rapid speed explosion requires the ability to automate a large proportion of the hardware manufacturing process. However, this kind of automation may already be achieved by the time that AGI is developed.[15]

### 2.3.3. Intelligence Explosion

Third, there could be an *intelligence explosion*, in which one AGI figures out how to create a qualitatively smarter AGI, and that AGI uses its increased intelligence to create still more intelligent AGIs, and so on,[16] such that the intelligence of humankind is quickly left far behind and the machines achieve dominance (Good 1965; Chalmers 2010; Muehlhauser and Salamon 2012; Loosemore and Goertzel 2012).[17]

Yudkowsky (2008a, 2008b) argues that an intelligence explosion is likely. So far, natural selection has been improving human intelligence, and human intelligence has to some extent been able to improve itself. However, the core process by which natural selection improves humanity has been essentially unchanged, and humans have been unable to deeply affect the cognitive algorithms which produce their own intelligence. Yudkowsky suggests that if a mind became capable of directly editing itself, this could spark a rapid increase in intelligence, as the actual process causing increases in intelligence could itself be improved upon. (This requires that there exist powerful improvements which, when implemented, considerably increase the rate at which such minds can improve themselves.)

Hall (2008) argues that, based on standard economic considerations, it would not make sense for an AGI to focus its resources on solitary self-improvement. Rather, in

15. Loosemore and Goertzel (2012) also suggest that current companies carrying out research and development are more constrained by a lack of capable researchers than by the ability to carry out physical experiments.

16. Most accounts of this scenario do not give exact definitions for "intelligence" or explain what a "superintelligent" AGI would be like, instead using informal characterizations such as "a machine that can surpass the intellectual activities of any man however clever" (Good 1965) or "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" (Bostrom 1998). Yudkowsky (2008a) defines intelligence in relation to "optimization power," the ability to reliably hit small targets in large search spaces, such as by finding the *a priori* exceedingly unlikely organization of atoms which makes up a car. A more mathematical definition of machine intelligence is offered by Legg and Hutter (2007). Sotala (2012) discusses some of the functional routes to actually achieving superintelligence.

17. One example of an AGI framework designed specifically for repeated self-improvement is offered by Schmidhuber (2009).

order not to be left behind by society at large, it should focus its resources on doing the things that it is good at and trade for the things it is not good at. However, once there exists a community of AGIs that can trade with one another, this community could collectively undergo rapid improvement and leave humans behind.

A number of formal growth models have been developed which are relevant to predicting the speed of a takeoff; an overview of these can be found in Sandberg's (2010) paper. Many of them suggest rapid growth. For instance, Hanson (1998) suggests that AGI might lead to the economy doubling in months rather than years. However, Hanson is skeptical about whether this would prove a major risk to humanity, and considers it mainly an economic transition similar to the Industrial Revolution.

To some extent, the soft/hard takeoff distinction may be a false dichotomy: a takeoff may be soft for a while, and then become hard. Two of the main factors influencing the speed of a takeoff are the pace at which computing hardware is developed and the ease of modifying minds (Sotala 2012). This allows for scenarios in which AGI is developed and there seems to be a soft takeoff for, say, the initial ten years, causing a false sense of security until a breakthrough in hardware development causes a hard takeoff.

Another factor that might cause a false sense of security is the possibility that AGIs can be developed by a combination of insights from humans and AGIs themselves. As AGIs become more intelligent and it becomes possible to automate portions of the development effort, those parts accelerate and the parts requiring human effort become bottlenecks. Reducing the amount of human insight required could dramatically accelerate the speed of improvement. Halving the amount of human involvement required might at most double the speed of development, possibly giving an impression of relative safety, but going from 50% human insight required to 1% human insight required could cause the development to become ninety-nine times faster.[18]

From a safety viewpoint, the conservative assumption is to presume the worst (Yudkowsky 2001). Yudkowsky argues that the worst outcome would be a hard takeoff, as it would give us the least time to prepare and correct errors. On the other hand, it can also be argued that a soft takeoff would be just as bad, as it would allow the creation of multiple competing AGIs, allowing the AGIs that were the least burdened with goals such as "respect human values" to prevail. We would ideally like a solution, or a combination of solutions, which would work effectively for both a soft and a hard takeoff.

18. The relationship in question is similar to that described by Amdahl's (1967) law.

## 3. Societal Proposals

The notion of catastrophic AGI risk is not new, and this concern was expressed by early thinkers in the field (Butler 1863; Turing 1951; Wiener 1960; Good 1965). Hence, there have also been many proposals concerning what to do about it. The proposals we survey are neither exhaustive nor mutually exclusive: the best way of achieving a desirable outcome may involve pursuing several proposals simultaneously.

Proposals can be divided into three general categories: proposals for societal action, design proposals for external constraints on AGI behavior, and design recommendations for internal constraints on AGI behavior. In this section we briefly survey societal proposals. These include doing nothing, integrating AGIs with society, regulating research, merging with machines, and relinquishing research into AGI.

### 3.1. Do Nothing

#### 3.1.1. AI Is Too Distant to Be Worth Our Attention

One response is that, although AGI is possible in principle, there is no reason to expect it in the near future. Typically, this response arises from the belief that, although there have been great strides in narrow AI, researchers are still very far from understanding how to build AGI. Distinguished computer scientists such as Gordon Bell and Gordon Moore, as well as cognitive scientists such as Douglas Hofstadter and Steven Pinker, have expressed the opinion that the advent of AGI is remote (*IEEE Spectrum* 2008). Davis (2012) reviews some of the ways in which computers are still far from human capabilities. Bringsjord and Bringsjord (2012) even claim that a belief in AGI this century is fideistic, appropriate within the realm of religion but not within science or engineering.

Some writers also actively criticize any discussion of AGI risk in the first place. The philosopher Alfred Nordmann (2007, 2008) holds the view that ethical concern is a scarce resource, not to be wasted on unlikely future scenarios such as AGI. Likewise, Dennett (2012) considers AGI risk an "imprudent pastime" because it distracts our attention from more immediate threats.

Others think that AGI is far off and not yet a major concern, but admit that it might be valuable to give the issue some attention. A presidential panel of the Association for the Advancement of Artificial Intelligence considering the long-term future of AI concluded that there was overall skepticism about AGI risk, but that additional research into the topic and related subjects would be valuable (Horvitz and Selman 2009). Posner (2004) writes that dedicated efforts for addressing the problem can wait, but that we should gather more information about the problem in the meanwhile.

Potential negative consequences of AGI are enormous, ranging from economic instability to human extinction. "Do nothing" could be a reasonable course of action if near-term AGI seemed extremely unlikely, if it seemed too early for any proposals to be effective in reducing risk, or if those proposals seemed too expensive to implement.

As a comparison, asteroid impact prevention is generally considered a topic worth studying, even though the probability of a civilization-threatening asteroid impact in the near future is not considered high. Napier (2008) discusses several ways of estimating the frequency of such impacts. Many models produce a rate of one civilization-threatening impact per five hundred thousand or more years, though some models suggest that rates of one such impact per hundred thousand years cannot be excluded.

An estimate of one impact per hundred thousand years would suggest less than a 0.1% chance of a civilization-threatening impact within the next hundred years. The probability of AGI being developed within the same period seems considerably higher (Muehlhauser and Salamon 2012), and there is likewise a reasonable chance of a hard takeoff after it has been developed (Yudkowsky 2008a, 2008b), suggesting that the topic is at the very least worth studying. Even without a hard takeoff, society is becoming increasingly automated, and even narrow AI is starting to require ethical guidelines (Wallach and Allen 2009; Arkin 2009).

We know neither which fields of science will be needed nor how much progress in them will be necessary for safe AGI. If much progress is needed and we believe effective progress to be possible this early on, it becomes reasonable to start studying the topic even before AGI is near. Muehlhauser and Helm (2012) suggest that, for one safe AGI approach alone (value learning, discussed further in section 5), efforts by AGI researchers, economists, mathematicians, and philosophers may be needed. Safe AI may require the solutions for some of these problems to come well before AGI is developed.

### 3.1.2. Little Risk, No Action Needed

Some authors accept that a form of AGI will probably be developed but do not consider autonomous AGI to be a risk, or consider the possible negative consequences acceptable. Bryson and Kime (1998) argue that, although AGI will require us to consider ethical and social dangers, the dangers will be no worse than those of other technologies. Whitby (1996) writes that there has historically been no consistent trend of the most intelligent people acquiring the most authority, and that computers will augment humans rather than replace them. Whitby and Oliver (2000) further argue that AGIs will not have any particular motivation to act against us. Jenkins (2003) agrees with these points to the extent of saying that a machine will only act against humans if it is programmed to value itself over humans, although she does find AGI to be a real concern.

Another kind of "no action needed" response argues that AGI development will take a long time (Brooks 2008), implying that there will be plenty of time to deal with the issue later on. This can also be taken as an argument for later efforts being more effective, as they will be better tuned to AGI as it develops.

Others argue that superintelligence will not be possible at all.[19] McDermott (2012) points out that there are no good examples of algorithms which could be improved upon *indefinitely*. Deutsch (2011) argues that there will never be superintelligent AGIs, because human minds are already universal reasoners, and computers can at best speed up the experimental work that is required for testing and fine-tuning theories. He also suggests that even as the speed of technological development increases, so will our ability to deal with change. M. Anderson (2010) likewise suggests that the inherent unpredictability of the world will place upper limits on an entity's effective intelligence.

Heylighen (2012) argues that a single, stand-alone computer is exceedingly unlikely to become superintelligent, and that individual intelligences are always outmatched by the distributed intelligence found in social systems of many minds. Superintelligence will be achieved by building systems that integrate and improve the "Global Brain," the collective intelligence of everyone on Earth. Heylighen does acknowledge that this kind of a transition will pose its own challenges, but not of the kind usually evoked in discussions of AGI risk.

The idea of AGIs not having a motivation to act against humans is intuitively appealing, but there seem to be strong theoretical arguments against it. As mentioned earlier, Omohundro (2007, 2008) and Bostrom (2012) argue that self-replication and the acquisition of resources are useful in the pursuit of many different kinds of goals, and that many types of AI systems will therefore exhibit tendencies toward behaviors such as breaking into other machines, self-replicating, and acquiring resources without regard for anyone else's safety. The right design might make it possible to partially work around these behaviors (Shulman 2010a; Wang 2012), but they still need to be taken into account. Furthermore, we might not foresee all the complex interactions of different AGI mechanisms in the systems that we build, and they may end up with very different goals than the ones we intended (Yudkowsky 2008a, 2011; Ring and Orseau 2011; Dewey 2011).

Can AGIs become superintelligent? First, we note that AGIs do not necessarily need to be much more intelligent than humans in order to be dangerous. AGIs already enjoy

19. The opposite argument is that superior intelligence will inevitably lead to more moral behavior. Some of the arguments related to this position are discussed in the context of evolutionary invariants (section 5.3.1), although the authors advocating the use of evolutionary invariants do believe AGI risk to be worth our concern.

advantages such as the ability to rapidly expand their population by having themselves copied (Hanson 1994, 2008; Sotala 2012), which may confer on them considerable economic and political influence even if their intelligence were near the human level. A better-than-human ability to coordinate their actions, which AGIs of a similar design could plausibly have (Sotala 2012), might then be enough to tilt the odds in their favor.

Another consideration is that AGIs do not necessarily need to be qualitatively more intelligent than humans in order to outperform humans. An AGI that merely thought twice as fast as any single human could still defeat him at intellectual tasks that had a time constraint, all else equal. Here an "intellectual" task should be interpreted broadly to refer not only to "book smarts" but to any task that animals cannot perform due to their mental limitations—including tasks involving social skills (Yudkowsky 2008a). Straightforward improvements in computing power could provide AGIs with a considerable advantage in speed (Solomonoff 1985; Yudkowsky 1996; Chalmers 2010; Hutter 2012; Sotala 2012), which the AGI could then use to study and accumulate experiences that improved its skills.

As for Heylighen's (2012) Global Brain argument, there does not seem to be a reason to presume that powerful AGIs could not be geographically distributed, or that they couldn't seize control of much of the Internet. Even if individual minds were not very smart and needed a society to make progress, for minds that are capable of copying themselves and communicating perfectly with each other, individual instances of the mind might be better understood as parts of a whole than as separate individuals (Sotala 2012; Sotala and Valpola 2012). In general, the distinction between an individual and a community might not be meaningful for AGIs (Goertzel and Pitt 2012). If there were enough AGIs, they might be able to form a community sufficient to take control of the rest of the Earth. Heylighen (2007) himself has argued that many of the features of the Internet are virtually identical to the mechanisms used by the human brain. If the AGI is not carefully controlled, it might end up in a position where it made up the majority of the "Global Brain" and could undertake actions which the remaining parts of the organism did not agree with.

### 3.1.3. Let Them Kill Us

Dietrich (2007) argues that humanity frequently harms other species, in part because we have evolved to engage in behaviors such as child abuse, sexism, rape, and racism. Therefore, human extinction would not matter, as long as the machines implemented only the positive aspects of humanity.

De Garis (2005) suggests that AGIs destroying humanity might not matter. He writes that on a cosmic scale, with hundreds of billions of stars in our galaxy alone,

the survival of the inhabitants of a single planet is irrelevant. As AGIs would be more intelligent than us in every way, it would be better if they replaced humanity.

AGIs being more intelligent and therefore more valuable than humans equates intelligence with value, but Bostrom (2004) suggests ways by which a civilization of highly intelligent entities might lack things which we thought to have value. For example, such entities might not be conscious in the first place. Alternatively, there are many things which we consider valuable for their own sake, such as humor, love, game-playing, art, sex, dancing, social conversation, philosophy, literature, scientific discovery, food and drink, friendship, parenting, and sport. We value these due to the fact that we have dispositions and preferences which have been evolutionarily adaptive in the past, but for a future civilization few or none of them might be, creating a world with very little of value.

#### 3.1.4. "Do Nothing" Proposals: Our View

As discussed above, completely ignoring the possibility of AGI risk at this stage would seem to require a confident belief in at least one of the following propositions:

1. AGI is very remote.
2. There is no major risk from AGI even if it is created.
3. Very little effective work can be done at this stage.
4. AGIs destroying humanity would not matter.

In the beginning of this paper, we mentioned several experts who considered it plausible that AGI might be created in the next twenty to one hundred years; in this section we have covered experts who disagree.

In general, there is a great deal of disagreement among people who have made AGI predictions, and no clear consensus even among experts in the field of artificial intelligence. The lack of expert agreement suggests that expertise in the field does not contribute to an ability to make reliable predictions (Armstrong and Sotala 2012).[20] If the judgment of experts is not reliable, then, probably, neither is anyone else's. This suggests that it is unjustified to be highly certain of AGI being near, but also of it *not* being near. We thus consider it unreasonable to have a confident belief in the first proposition.

20. Armstrong and Sotala (2012) point out that many of the task properties which have been found to be conducive for developing reliable and useful expertise are missing in AGI timeline forecasting. In particular, one of the most important factors is whether experts get rapid (preferably immediate) feedback, while a timeline prediction that is set many decades in the future might have been entirely forgotten by the time that its correctness could be evaluated.

The second proposition also seems questionable: as discussed in sections 2.1, 2.3, and 3.1.2, AGIs seem very likely to obtain great power, possibly very quickly. Furthermore, as discussed in section 2.2, the complexity and fragility of value theses imply that it could be very difficult to create AGIs which would not cause immense amounts of damage if they had enough power.

It also does not seem like it is too early to work on the problem: as we summarize in section 6, there seem to be a number of promising research directions which can already be pursued. We also agree with Yudkowsky (2008a), who points out that research on the philosophical and technical requirements of safe AGI might show that broad classes of possible AGI architectures are fundamentally unsafe, suggesting that such architectures should be avoided. If this is the case, it seems better to have that knowledge as early as possible, before there has been a great deal of investment into unsafe AGI designs.

In response to the suggestion that humanity being destroyed would not matter, we certainly agree that there is much to be improved in today's humanity, and that our future descendants might have very little resemblance to ourselves. Regardless, we think that much about today's humans is valuable and worth preserving, and that we should be able to preserve it without involving the death of present humans.

### 3.2. Integrate With Society

Integration proposals hold that AGI might be created in the next several decades, and that there are indeed risks involved. These proposals argue that the best way to deal with the problem is to make sure that our societal structures are equipped to handle AGIs once they are created.

There has been some initial work toward integrating AGIs with existing legal and social frameworks, such as considering questions of their legal position (Lehman-Wilzig 1981; Solum 1992; Asaro 2007; Bryson 2010; Weng, Chen, and Sun 2008, 2009; Hallevy 2010) and moral rights (Sullins 2005, 2006; Bryson 2010; Levy 2009; Warwick 2010; Gunkel 2012).

#### 3.2.1. Legal and Economic Controls

Hanson (2012) writes that the values of older and younger generations have often been in conflict with each other, and he compares this to a conflict between humans and AGIs. He believes that the best way to control AGI risk is to create a legal framework such that it is in the interest of both humans and AGIs to uphold it. Hanson (2009) suggests that if the best way for AGIs to get what they want is via mutually agreeable exchanges, then humans would need to care less about what the AGIs wanted. According to him, we should be primarily concerned with ensuring that the AGIs will be law-abiding enough to respect our property rights. Miller (2012) summarizes Hanson's argument, and the

idea that humanity could be content with a small fraction of the world's overall wealth and let the AGIs have the rest. Analogously, many Westerners past retirement age are wealthy, despite no longer contributing to production and despite the fact that others could join together and take their wealth if they really wanted to.

Hall (2007a) also says that we should ensure that the interactions between ourselves and machines are economic, "based on universal rules of property and reciprocity." Moravec (1999) likewise writes that governmental controls should be used to ensure that humans benefit from AGIs. Without government intervention, humans would be squeezed out of existence by more efficient robots, but taxation could be used to support human populations for a long time. He also recommends laws which would require any AGIs to incorporate programming that made them safe and subservient to human desires. Sandberg (2001) writes that relying only on legal and economic controls would be problematic, but that a strategy which also incorporated them in addition to other approaches would be more robust than a strategy which did not.

However, even if AGIs were integrated with human institutions, it does not guarantee that human values would survive. If humans were reduced to a position of negligible power, AGIs might not have any reason to keep us around.

Economic arguments, such as the principle of comparative advantage, are sometimes invoked to argue that AGI would find it more beneficial to trade with us than to do us harm. However, technological progress can drive the wages of workers below the level needed for survival (Clark 2007; Freeman 2008; Brynjolfsson and McAfee 2011; Miller 2012), and there is already a possible threat of technological unemployment (Brynjolfsson and McAfee 2011). AGIs keeping humans around due to gains from trade implicitly presumes that they would not have the will or the opportunity to simply eliminate humans in order to replace them with a better trading partner, and then trade with the new partner instead.

Humans already eliminate species with low economic value in order to make room for more humans, such as when clearing a forest in order to build new homes. Clark uses the example of horses in Britain: their population peaked in 1901, with 3.25 million horses doing work such as plowing fields, hauling wagons and carriages short distances, and carrying armies into battle. The internal combustion engine replaced so many of them that by 1924 there were fewer than two million. Clark writes:

> There was always a wage at which all these horses could have remained employed. But that wage was so low that it did not pay for their feed, and it certainly did not pay enough to breed fresh generations of horses to replace them. Horses were thus an early casualty of industrialization. (Clark 2007)

There are also ways to harm humans while still respecting their property rights, such as by manipulating them into making bad decisions, or selling them addictive substances.

If AGIs were sufficiently smarter than humans, humans could be tricked into making a series of trades that respected their property rights but left them with negligible assets and caused considerable damage to their well-being.

A related issue is that AGIs might become more capable of changing our values than we are capable of changing AGI values. Mass media already convey values that have a negative impact on human well-being, such as idealization of rare body types, which causes dissatisfaction among people who do not have those kinds of bodies (Groesz, Levine, and Murnen 2001; Agliata and Tantleff-Dunn 2004). AGIs with a deep understanding of human psychology could engineer the spread of values which shifted more power to them, regardless of their effect on human well-being.

Yet another problem is ensuring that the AGIs have indeed adopted the right values. Making intelligent beings adopt specific values is a difficult process which often fails. There could be an AGI with the wrong goals that would pretend to behave correctly in society throughout the whole socialization process. AGIs could conceivably preserve and conceal their goals far better than humans could.

Society does not know of any methods which would reliably instill our chosen values in *human* minds, despite a long history of trying to develop them. Our attempts to make AGIs adopt human values would be hampered by our lack of experience and understanding of the AGI's thought processes, with even tried-and-true methods for instilling positive values in humans possibly being ineffective. The limited success that we do have with humans is often backed up by various incentives as well as threats of punishment, both of which might fail in the case of an AGI developing to become vastly more powerful than us.

Additionally, the values which a being is likely to adopt, or is even capable of adopting, will depend on its mental architecture. We will demonstrate these claims with examples from humans, who are not blank slates on whom arbitrary values can be imposed with the right education (Pinker 2002). Although the challenge of instilling specific values in humans is very different from the challenge of instilling them in AGIs, our examples are meant to demonstrate the fact that the existing properties of a mind will affect the process of acquiring values. Just as it is difficult to make humans permanently adopt some kinds of values, the kind of mental architecture that an AGI has will affect its inclination to adopt various values.

Psychopathy is a risk factor for violence, and psychopathic criminals are much more likely to reoffend than nonpsychopaths (Hare et al. 2000). Harris and Rice (2006) argue that therapy for psychopaths is ineffective[21] and may even make them more dangerous,

21. Salekin (2010) offers a more optimistic opinion.

as they use their improved social skills to manipulate others more effectively. Furthermore, "cult brainwashing" is generally ineffective and most cult members will eventually leave (Anthony and Robbins 2004); and large-scale social engineering efforts often face widespread resistance, even in dictatorships with few scruples about which methods to use (Scott 1998, chap. 6–7). Thus, while one can try to make humans adopt values, this will only work to the extent that the individuals in question are actually disposed toward adopting them.

#### 3.2.2. Foster Positive Values

Kurzweil (2005), considering the possible effects of many future technologies, notes that AGI may be a catastrophic risk. He generally supports regulation and partial relinquishment of dangerous technologies, as well as research into their defensive applications. However, he believes that with AGI this may be insufficient and that, at the present time, it may be infeasible to develop strategies that would guarantee safe AGI. He argues that machine intelligences will be tightly integrated into our society and that, for the time being, the best chance of avoiding AGI risk is to foster positive values in our society. This will increase the likelihood that any AGIs that are created will reflect such positive values.

One possible way of achieving such a goal is moral enhancement (Douglas 2008), the use of technology to instill people with better motives. Persson and Savulescu (2008, 2012) argue that, as technology improves, we become more capable of damaging humanity, and that we need to carry out moral enhancement in order to lessen our destructive impulses.

#### 3.2.3. "Integrate With Society" Proposals: Our View

Proposals to incorporate AGIs into society suffer from the issue that some AGIs may never adopt benevolent and cooperative values, no matter what the environment. Neither does the intelligence of the AGIs necessarily affect their values (Bostrom 2012). Sufficiently intelligent AGIs could certainly come to eventually understand human values, but humans can also come to understand others' values while continuing to disagree with them.

Thus, in order for these kinds of proposals to work, they need to incorporate strong enforcement mechanisms to keep non–safe AGIs in line and to prevent them from acquiring significant power. This requires an ability to create value-conforming AGIs in the first place, to implement the enforcement. Even a soft takeoff would eventually lead to AGIs wielding great power, so the enforcement could not be left to just humans

or narrow AIs.[22] In practice, this means that integration proposals must be combined with some proposal for internal constraints which is capable of reliably creating value-conforming AGIs. Integration proposals also require there to be a soft takeoff in order to work, as having a small group of AGIs which rapidly acquired enough power to take control of the world would prevent any gradual integration schemes from working.

Therefore, because any effective integration strategy would require creating safe AGIs, and the right safe AGI design could lead to a positive outcome even if there were a hard takeoff, we believe that it is currently better to focus on proposals which are aimed at furthering the creation of safe AGIs.

### 3.3. Regulate Research

Integrating AGIs into society may require explicit regulation. Calls for regulation are often agnostic about long-term outcomes but nonetheless recommend caution as a reasonable approach. For example, Hibbard (2005b) calls for international regulation to ensure that AGIs will value the long-term well-being of humans, but does not go into much detail. Daley (2011) calls for a government panel for AGI issues. Hughes (2001) argues that AGI should be regulated using the same mechanisms as previous technologies, creating state agencies responsible for the task and fostering global cooperation in the regulation effort.

Current mainstream academic opinion does not consider AGI a serious threat (Horvitz and Selman 2009), so AGI regulation seems unlikely in the near future. On the other hand, many AI systems are becoming increasingly autonomous, and a number of authors are arguing that even narrow-AI applications should be equipped with an understanding of ethics (Wallach and Allen 2009). Currently there are calls to regulate AI in the form of high-frequency trading (Sobolewski 2012), and AI applications that have a major impact on society might become increasingly regulated. At the same time, legislation has a well-known tendency to lag behind technology, and regulating AI applications will probably not translate into regulating basic research into AGI.

#### 3.3.1. Review Boards

Yampolskiy and Fox (2012) note that university research programs in the social and medical sciences are overseen by institutional review boards. They propose setting up analogous review boards to evaluate potential AGI research. Research that was found to be AGI related would be restricted with measures ranging from supervision and funding

22. For proposals which suggest that humans could use technology to remain competitive with AGIs and thus prevent them from acquiring excessive amounts of power, see section 3.4.

limits to partial or complete bans. At the same time, research focusing on safety measures would be encouraged.

Posner (2004, p. 221) suggests the enactment of a law which would require scientific research projects in dangerous areas to be reviewed by a federal catastrophic risks assessment board, and forbidden if the board found that the project would create an undue risk to human survival.

G. S. Wilson (forthcoming) makes possibly the most detailed AGI regulation proposal so far, recommending a new international treaty where a body of experts would determine whether there was a "reasonable level of concern" about AGI or some other possibly dangerous research. States would be required to regulate research or even temporarily prohibit it once experts agreed upon there being such a level of concern. He also suggests a number of other safeguards built into the treaty, such as the creation of ethical oversight organizations for researchers, mechanisms for monitoring abuses of dangerous technologies, and an oversight mechanism for scientific publications.

### 3.3.2. Encourage Research into Safe AGI

In contrast, McGinnis (2010) argues that the government should not attempt to regulate AGI development. Rather, it should concentrate on providing funding for research projects intended to create safe AGI.

Goertzel and Pitt (2012) argue for an open-source approach to safe AGI development instead of regulation. Hibbard (2008) has likewise suggested developing AGI via open-source methods, but not as an alternative to regulation.

Legg (2009) proposes funding safe AGI research via an organization that takes a venture capitalist approach to funding research teams, backing promising groups and cutting funding to any teams that fail to make significant progress. The focus of the funding would be to make AGI as safe as possible.

### 3.3.3. Differential Technological Progress

Both review boards and government funding could be used to implement "differential intellectual progress":

> Differential intellectual progress consists in prioritizing risk-reducing intellectual progress over risk-increasing intellectual progress. As applied to AI risks in particular, a plan of differential intellectual progress would recommend that our progress on the scientific, philosophical, and technological problems of AI safety outpace our progress on the problems of AI capability such that we develop safe superhuman AIs before we develop (arbitrary) superhuman AIs. (Muehlhauser and Salamon 2012)

Examples of research questions that could constitute philosophical or scientific progress in safety can be found in later sections of this paper—for instance, the usefulness of different internal constraints on ensuring safe behavior, or ways of making AGIs reliably adopt human values as they learn what those values are like.

Earlier, Bostrom (2002) used the term "differential technological progress" to refer to differential intellectual progress in technological development. Bostrom defined differential technological progress as "trying to retard the implementation of dangerous technologies and accelerate implementation of beneficial technologies, especially those that ameliorate the hazards posed by other technologies."

One issue with differential technological progress is that we do not know what kind of progress should be accelerated and what should be retarded. For example, a more advanced communication infrastructure could make AGIs more dangerous, as there would be more networked machines that could be accessed via the Internet. Alternatively, it could be that the world will already be so networked that AGIs will be a major threat anyway, and further advances will make the networks more resilient to attack. Similarly, it can be argued that AGI development is dangerous for as long as we have yet to solve the philosophical problems related to safe AGI design and do not know which AGI architectures are safe to pursue (Yudkowsky 2008a). But it can also be argued that we should invest in AGI development now, when the related tools and hardware are still primitive enough that progress will be slow and gradual (Goertzel and Pitt 2012).

#### 3.3.4. International Mass Surveillance

For AGI regulation to work, it needs to be enacted on a global scale. This requires solving both the problem of effectively enforcing regulation within a country and the problem of getting many different nations to all agree on the need for regulation.

Shulman (2009) discusses various factors influencing the difficulty of AGI arms control. He notes that AGI technology itself might make international cooperation more feasible. If narrow AIs and early-stage AGIs were used to analyze the information obtained from wide-scale mass surveillance and wiretapping, this might make it easier to ensure that nobody was developing more advanced AGI designs.

Shulman (2010b) similarly notes that machine intelligences could be used to enforce treaties between nations. They could also act as trustworthy inspectors which would be restricted to communicating only information about treaty violations, thus not endangering state secrets even if they were allowed unlimited access to them. This could help establish a "singleton" (Bostrom 2003b) regulatory regimen capable of effectively enforcing international regulation, including AGI-related treaties. Goertzel and Pitt (2012) also discuss the possibility of having a network of AGIs monitoring the world in

order to police other AGIs and to prevent any of them from suddenly obtaining excessive power.

Another proposal for international mass surveillance is to build an "AGI Nanny" (Goertzel 2012b; Goertzel and Pitt 2012), a proposal discussed in section 5.4.

Large-scale surveillance efforts are ethically problematic and face major political resistance, and it seems unlikely that current political opinion would support the creation of a far-reaching surveillance network for the sake of AGI risk alone. The extent to which such extremes would be necessary depends on exactly how easy it would be to develop AGI in secret. Although several authors make the point that AGI is much easier to develop unnoticed than something like nuclear weapons (McGinnis 2010; Miller 2012), cutting-edge high-tech research does tend to require major investments which might plausibly be detected even by less elaborate surveillance efforts.

To the extent that surveillance does turn out to be necessary, there is already a strong trend toward a "surveillance society" with increasing amounts of information about people being collected and recorded in various databases (Wood and Ball 2006). As a reaction to the increased surveillance, Mann, Nolan, and Wellman (2003) propose to counter it with *sous*veillance—giving private individuals the ability to document their life and subject the authorities to surveillance in order to protect civil liberties. This is similar to the proposals of Brin (1998), who argues that technological progress might eventually lead to a "transparent society," where we will need to redesign our societal institutions in a way that allows us to maintain some of our privacy despite omnipresent surveillance. Miller (2012) notes that intelligence agencies are already making major investments in AI-assisted analysis of surveillance data.

If social and technological developments independently create an environment where large-scale surveillance or sousveillance is commonplace, it might be possible to take advantage of those developments in order to police AGI risk.[23] Walker (2008) argues that in order for mass surveillance to become effective, it must be designed in such a way that it will not excessively violate people's privacy, for otherwise the system will face widespread sabotage.[24] Even under such conditions, there is no clear way to define

23. An added benefit would be that this could also help avoid other kinds of existential risk, such as the intentional creation of dangerous new diseases.

24. Walker also suggests that surveillance systems could be designed to automatically edit out privacy-endangering details (such as pictures of people) from the data that they transmit, while leaving in details which might help in revealing dangerous ploys (such as pictures of bombs). However, this seems impossible to implement effectively, as research has found ways to extract personally identifying information and details from a wide variety of supposedly anonymous datasets (Sweeney 1997; Felten and Schneider 2000; Narayanan and Shmatikov 2008, 2009a; Golle and Partridge 2009; Calandrino, Clarkson, and Felten 2011; Narayanan et al. 2012). Narayanan and Shmatikov (2009b) even go as far as to

what counts as dangerous AGI. Goertzel and Pitt (2012) point out that there is no clear division between narrow AI and AGI, and attempts to establish such criteria have failed. They argue that since AGI has a nebulous definition, obvious wide-ranging economic benefits, and potentially rich penetration into multiple industry sectors, it is unlikely to be regulated due to speculative long-term risks.

AGI regulation requires global cooperation, as the noncooperation of even a single nation might lead to catastrophe. Historically, achieving global cooperation on tasks such as nuclear disarmament and climate change has been very difficult. As with nuclear weapons, AGI could give an immense economic and military advantage to the country that develops it first, in which case limiting AGI research might even give other countries an incentive to develop AGI faster (Cade 1966; de Garis 2005; McGinnis 2010; Miller 2012).

To be effective, regulation also needs to enjoy support among those being regulated. If developers working in AGI-related fields only follow the letter of the law, while privately viewing all regulations as annoying hindrances, and fears about AGI as overblown, the regulations may prove ineffective. Thus, it might not be enough to convince governments of the need for regulation; the much larger group of people working in the appropriate fields may also need to be convinced.

While Shulman (2009) argues that the unprecedentedly destabilizing effect of AGI could be a cause for world leaders to cooperate more than usual, the opposite argument can be made as well. Gubrud (1997) argues that increased automation could make countries more self-reliant, and international cooperation considerably more difficult. AGI technology is also much harder to detect than, for example, nuclear technology is—nuclear weapons require a substantial infrastructure to develop, while AGI needs much less (McGinnis 2010; Miller 2012).

Miller (2012) even suggests that the mere possibility of a rival being close to developing AGI might, if taken seriously, trigger a nuclear war. The nation that was losing the AGI race might think that being the first to develop AGI was sufficiently valuable that it was worth launching a first strike for, even if it would lose most of its own population in the retaliatory attack. He further argues that, although it would be in the interest of every nation to try to avoid such an outcome, the ease of secretly pursuing an AGI development program undetected, in violation of treaty, could cause most nations to violate the treaty.

---

state that "the false dichotomy between personally identifiable and non-personally identifiable information should disappear from privacy policies, laws, etc. Any aspect of an individual's . . . personality can be used for de-anonymization, and this reality should be recognized by the relevant legislation and corporate privacy policies."

Miller also points out that the potential for an AGI arms race exists not only between nations, but between corporations as well. He notes that the more AGI developers there are, the more likely it is that they will all take more risks, with each AGI developer reasoning that if they don't take this risk, somebody else might take that risk first.

Goertzel and Pitt (2012) suggest that for regulation to be enacted, there might need to be an "AGI Sputnik"—a technological achievement that makes the possibility of AGI evident to the public and policy makers. They note that after such a moment, it might not take very long for full human-level AGI to be developed, while the negotiations required to enact new kinds of arms control treaties would take considerably longer.

So far, the discussion has assumed that regulation would be carried out effectively and in the pursuit of humanity's common interests, but actual legislation is strongly affected by lobbying and the desires of interest groups (Olson 1982; Mueller 2003, chap. 22). Many established interest groups would have an economic interest in either furthering or retarding AGI development, rendering the success of regulation uncertain.

#### 3.3.5. "Regulate Research" Proposals: Our View

Although there seem to be great difficulties involved with regulation, there also remains the fact that many technologies have been successfully subjected to international regulation. Even if one were skeptical about the chances of effective regulation, an AGI arms race seems to be one of the worst possible scenarios, one which should be avoided if at all possible. We are therefore generally supportive of regulation, though the most effective regulatory approach remains unclear.

### 3.4. Enhance Human Capabilities

While regulation approaches attempt to limit the kinds of AGIs that will be created, enhancement approaches attempt to give humanity and AGIs a level playing field. In principle, gains in AGI capability would not be a problem if humans could improve themselves to the same level.

Alternatively, human capabilities could be improved in order to obtain a more general capability to deal with difficult problems. Verdoux (2010, 2011) suggests that cognitive enhancement could help in transforming previously incomprehensible mysteries into tractable problems, and Verdoux (2010) particularly highlights the possibility of cognitive enhancement helping to deal with the problems posed by existential risks. One problem with such approaches is that increasing humanity's capability for solving problems will also make it easier to develop dangerous technologies. It is possible that cognitive enhancement should be combined with moral enhancement, in order to help foster the kind of cooperation that would help avoid the risks of technology (Persson and Savulescu 2008, 2012).

Moravec (1988, 1999) proposes that humans could keep up with AGIs via "mind uploading," a process of transferring the information in human brains to computer systems so that human minds could run on a computer substrate. This technology may arrive during a similar timeframe as AGI (Kurzweil 2005; Sandberg and Bostrom 2008; Hayworth 2012; Koene 2012b; Sotala and Valpola 2012; Cattell and Parker 2012; Sandberg 2012). However, Moravec argues that mind uploading would come after AGIs, and that unless the uploads would transform themselves to become radically nonhuman, they would be weaker and less competitive than AGIs that were native to a digital environment (Moravec 1992, 1999). For these reasons,Warwick (1998) also expresses doubt about the usefulness of mind uploading.[25]

Kurzweil (2005) posits an evolution that will start with brain-computer interfaces, then proceed to using brain-embedded nanobots to enhance our intelligence, and finally lead to full uploading and radical intelligence enhancement. Koene (2012a) criticizes plans to create safe AGIs and considers uploading both a more feasible and a more reliable approach.

Similar proposals have also been made without explicitly mentioning mind uploading. Cade (1966) speculates on the option of gradually merging with machines by replacing body parts with mechanical components. Turney (1991) proposes linking AGIs directly to human brains so that the two meld together into one entity, and Warwick (1998, 2003) notes that cyborgization could be used to enhance humans.

Mind uploading might also be used to make human value systems more accessible and easy to learn for AGIs, such as by having an AGI extrapolate the upload's goals directly from its brain, with the upload providing feedback.

### 3.4.1. Would We Remain Human?

Uploading might destroy parts of humanity that we value (Joy 2000; de Garis 2005). De Garis (2005) argues that a computer could have far more processing power than a human brain, making it pointless to merge computers and humans. The biological component of the resulting hybrid would be insignificant compared to the electronic component, creating a mind that was negligibly different from a "pure" AGI. Kurzweil (2001) makes the same argument, saying that although he supports intelligence enhancement by di-

25. Some uploading approaches also raise questions of personal identity, whether the upload would still be the same person as the original (Moravec 1988; Chalmers 2010; Walker 2011; Bamford 2012; Blackmore 2012; Goertzel 2012c; Hauskeller 2012; Hopkins 2012; Swan and Howard 2012), and whether they would be conscious in the first place (Searle 1992; Chalmers 1996, pp. 247–275; Kurzweil 2002; Agar 2011; Hauskeller 2012). However, these concerns are not necessarily very relevant for AGI risk considerations, as a population of uploads working to protect against AGIs would be helpful even if they lacked consciousness or were different individuals than the originals.

rectly connecting brains and computers, this would only keep pace with AGIs for a couple of additional decades.

The truth of this claim seems to depend on exactly how human brains are augmented. In principle, it seems possible to create a prosthetic extension of a human brain that uses the same basic architecture as the original brain and gradually integrates with it (Sotala and Valpola 2012). A human extending their intelligence using such a method might remain roughly human-like and maintain their original values. However, it could also be possible to connect brains with computer programs that are very unlike human brains, and which would substantially change the way the original brain worked. Even smaller differences could conceivably lead to the adoption of "cyborg values" distinct from ordinary human values (Warwick 2003).

Bostrom (2004) speculates that humans might outsource many of their skills to non-conscious external modules and would cease to experience anything as a result. The value-altering modules would provide substantial advantages to their users, to the point that they could outcompete uploaded minds who did not adopt the modules.

Uploading would also allow humans to make precise and deep modifications to their own minds, which carries with it dangers of a previously unencountered kind (Suber 2002).

### 3.4.2. Would Evolutionary Pressures Change Us?

A willingness to integrate value-altering modules is not the only way by which a population of uploads might come to have very different values from modern-day humans. This is not necessarily a bad, or even a very novel, development: the values of earlier generations have often been different from the values of later generations (Hanson 2012), and it might not be a problem if a civilization of uploads enjoyed very different things than a civilization of humans. Still, as there are possible outcomes that we would consider catastrophic, such as the loss of nearly all things that have intrinsic value for us (Bostrom 2004), it is worth reviewing some of the postulated changes in values.

For comprehensiveness, we will summarize all of the suggested effects that uploading might have on human values, even if they are not obviously negative. Readers may decide for themselves whether or not they consider any of these effects concerning.

Hanson (1994) argues that employers will want to copy uploads who are good workers, and that at least some uploads will consent to being copied in such a manner. He suggests that the resulting evolutionary dynamics would lead to an accelerated evolution of values. This would cause most of the upload population to evolve to be indifferent or favorable to the thought of being copied, to be indifferent toward being deleted as long as another copy of themselves remained, and to be relatively uninterested in having children "the traditional way" (as opposed to copying an already-existing mind). Al-

though Hanson's analysis uses the example of a worker-employer relationship, it should be noted that nations or families, or even single individuals, could also gain a competitive advantage by copying themselves, thus contributing to the strength of the evolutionary dynamic.

Similarly, Bostrom (2004) writes that much of human life's meaning depends on the enjoyment of things ranging from humor and love to literature and parenting. These capabilities were adaptive in our past, but in an upload environment they might cease to be such and gradually disappear entirely.

Shulman (2010b) likewise considers uploading-related evolutionary dynamics. He notes that there might be a strong pressure for uploads to make copies of themselves in such a way that individual copies would be ready to sacrifice themselves to aid the rest. This would favor a willingness to copy oneself, and a view of personal identity which did not consider the loss of a single copy to be death. Beings taking this point of view could then take advantage of economic benefits of continually creating and deleting vast numbers of minds depending on the conditions, favoring the existence of a large number of short-lived copies over a somewhat less efficient world of long-lived minds.

Finally, Sotala and Valpola (2012) consider the possibility of minds coalescing via artificial connections that linked several brains together in the same fashion as the two hemispheres of ordinary brains are linked together. If this were to happen, considerable benefits might accrue to those who were ready to coalesce with other minds. The ability to copy and share memories between minds might also blur distinctions between individual minds. In the end, most humans might cease to be individual, distinct people in any real sense of the word.

It has also been proposed that information security concerns could cause undesirable dynamics among uploads, with significant advantages accruing to those who could steal the computational resources of others and use them to create new copies of themselves. If one could seize control of the hardware that an upload was running on, it could be immediately replaced with a copy of a mind loyal to the attacker. It might even be possible to do this without being detected, if it was possible to steal enough of an upload's personal information to impersonate it.

An attack targeting a critical vulnerability in some commonly used piece of software might quickly hit a very large number of victims. As previously discussed in section 2.3.1, both theoretical arguments and actual cases of malware show that large numbers of machines on the Internet could be infected in a very short time (Staniford, Paxson, and Weaver 2002; Moore, Shannon, and Brown 2002; Moore et al. 2003). In a society of uploads, attacks such as these would be not only inconvenient, but potentially fatal. Eckersley and Sandberg (forthcoming) offer a preliminary analysis of information security in a world of uploads.

### 3.4.3. Would Uploading Help?

Even if the potential changes of values were deemed acceptable, it is unclear whether the technology for uploading could be developed before developing AGI. Uploading might require emulating the low-level details of a human brain with a high degree of precision, requiring large amounts of computing power (Sandberg and Bostrom 2008; Cattell and Parker 2012). In contrast, an AGI might be designed around high-level principles which have been chosen to be computationally cheap to implement on existing hardware architectures.

Yudkowsky (2008a) uses the analogy that it is much easier to figure out the principles of aerodynamic flight and then build a Boeing 747 than it is to take a living bird and "upgrade" it into a giant bird that can carry passengers, all while ensuring that the bird remains alive and healthy throughout the process. Likewise, it may be much easier to figure out the basic principles of intelligence and build AGIs than to upload existing minds.

On the other hand, one can also construct an analogy suggesting that it is easier to copy a thing's function than it is to understand how it works. If a person does not understand architecture but wants to build a sturdy house, it may be easier to create a replica of an existing house than it is to design an entirely new one that does not collapse.

Even if uploads were created first, they might not be able to harness all the advantages of digitality, as many of these advantages depend on minds being easy to modify (Sotala 2012), which human minds may not be. Uploads will be able to directly edit their source code as well as introduce simulated pharmaceutical and other interventions, and they could experiment on copies that are restored to an unmodified state if the modifications turn out to be unworkable (Shulman 2010b). Regardless, human brains did not evolve to be easy to modify, and it may be difficult to find a workable set of modifications which would drastically improve them.

In contrast, in order for an AGI programmed using traditional means to be manageable as a software project, it must be easy for the engineers to modify it.[26] Thus, even if uploading were developed before AGI, AGIs that were developed later might still be capable of becoming more powerful than uploads. However, existing uploads already enjoying some of the advantages of the newly-created AGIs would still make it easier for the uploads to control the AGIs, at least for a while.

Moravec (1992) notes that the human mind has evolved to function in an environment which is drastically different from a purely digital environment, and that the only

26. However, this might not be true for AGIs created using some alternate means, such as artificial life (Sullins 2005).

way to remain competitive with AGIs would be to transform into something that was very different from a human. This suggests that uploading might buy time for other approaches, but would be only a short-term solution in and of itself.

If uploading technology were developed before AGI, it could be used to upload a research team or other group and run them at a vastly accelerated rate as compared to the rest of humanity. This would give them a considerable amount of extra time for developing any of the other approaches. If this group were among the first to be successfully emulated and sped up, and if the speed-up would allow enough subjective time to pass before anyone else could implement their own version, they might also be able to avoid trading safety for speed. However, such a group might be able to wield tremendous power, so they would need to be extremely reliable and trustworthy.

#### 3.4.4. "Enhance Human Capabilities" Proposals: Our View

Of the various "enhance human capabilities" approaches, uploading proposals seem the most promising, as translating a human brain to a computer program would sidestep many of the constraints that come from modifying a physical system. For example, all relevant brain activity could be recorded for further analysis at an arbitrary level of detail, and any part of the brain could be instantly modified without a need for time-consuming and possibly dangerous invasive surgery. Uploaded brains could also be more easily upgraded to take full advantage of more powerful hardware, while humans whose brains were still partially biological would be bottlenecked by the speed of the biological component.

Uploading does have several problems: Uploading research might lead to AGI being created before the uploads, in the long term uploads might have unfavorable evolutionary dynamics, and it seems likely that there will eventually be AGIs which are capable of outperforming uploads in every field of competence. Uploads could also be untrustworthy even without evolutionary dynamics. At the same time, however, uploading research doesn't *necessarily* accelerate AGI research very much, the evolutionary dynamics might not be as bad as they seem at the moment, and the advantages gained from uploading might be enough to help control unsafe AGIs until safe ones could be developed. Methods could also be developed for increasing the trustworthiness of uploads (Shulman 2010b). Uploading might still turn out to be a useful tool for handling AGI risk.

### 3.5. Relinquish Technology

Not everyone believes that the risks involved in creating AGIs are acceptable. *Relinquishment* involves the abandonment of technological development that could lead to AGI. This is possibly the earliest proposed approach, with Butler (1863) writing that "war to the death should be instantly proclaimed" upon machines, for otherwise they

would end up destroying humans entirely. In a much-discussed article, Joy (2000) suggests that it might be necessary to relinquish at least some aspects of AGI research, as well as nanotechnology and genetics research.

Hughes (2001) criticizes AGI relinquishment, while Kurzweil (2005) criticizes broad relinquishment but supports the possibility of "fine-grained relinquishment," banning some dangerous aspects of technologies while allowing general work on them to proceed. In general, most writers reject proposals for broad relinquishment.

### 3.5.1. Outlaw AGI

Weng, Chen, and Sun (2009) write that the creation of AGIs would force society to shift from human-centric values to robot-human dual values. In order to avoid this, they consider the possibility of banning AGI. This could be done either permanently or until appropriate solutions are developed for mediating such a conflict of values.

McKibben (2003), writing mainly in the context of genetic engineering, suggests that AGI research should be stopped. He brings up the historical examples of China renouncing seafaring in the 1400s and Japan relinquishing firearms in the 1600s, as well as the more recent decisions to abandon DDT, CFCs, and genetically modified crops in Western countries. However, it should also be noted that Japan participated in World War II; that China now has a navy; that there are reasonable alternatives for DDT and CFCs, which probably do not exist for AGI; and that genetically modified crops are in wide use in the United States.

Hughes (2001) argues that attempts to outlaw a technology will only make the technology move to other countries. He also considers the historical relinquishment of biological weapons to be a bad example, for no country has relinquished peaceful biotechnological research such as the development of vaccines, nor would it be desirable to do so. With AGI, there would be no clear dividing line between safe and dangerous research.

De Garis (2005) believes that differences of opinion about whether to build AGIs will eventually lead to armed conflict, to the point of open warfare. Annas, Andrews, and Isasi (2002) have similarly argued that genetic engineering of humans would eventually lead to war between unmodified humans and the engineered "posthumans," and that cloning and inheritable modifications should therefore be banned. To the extent that one accepts their reasoning with regard to humans, it could also be interpreted to apply to AGIs.

### 3.5.2. Restrict Hardware

Berglas (2012) suggests not only stopping AGI research, but also outlawing the production of more powerful hardware. He argues that, since computers are already nearly powerful enough to host an AGI, we should reduce the power of new processors and

destroy existing ones.[27] Branwen (2012) argues that Moore's Law depends on the existence of a small number of expensive and centralized chip factories, making them easy targets for regulation and incapable of being developed in secret.

### 3.5.3. "Relinquish Technology" Proposals: Our View

Relinquishment proposals suffer from many of the same problems as regulation proposals, but to a greater extent. There is no historical precedent of general, multiuse technology similar to AGI being successfully relinquished for good, nor do there seem to be any theoretical reasons for believing that relinquishment proposals would work in the future. Therefore we do not consider them to be a viable class of proposals.

## 4. External AGI Constraints

Societal approaches involving regulation or research into safe AGI assume that proper AGI design can produce solutions to AGI risk. One category of such solutions is that of *external constraints*. These are restrictions that are imposed on an AGI from the outside and aim to limit its ability to do damage.

Several authors have argued that external constraints are unlikely to work with AGIs that are genuinely far more intelligent than us (Vinge 1993; Yudkowsky 2001, 2008a; Kurzweil 2005; Chalmers 2010; Armstrong, Sandberg, and Bostrom 2012). The consensus seems to be that external constraints might buy time when dealing with less advanced AGIs, but they are useless against truly superintelligent ones.

External constraints also limit the usefulness of an AGI, as a free-acting one could serve its creators more effectively. This reduces the probability of the universal implementation of external constraints on AGIs. AGIs might also be dangerous if they were confined or otherwise restricted. For further discussion of these points, see section 5.1.

### 4.1. AGI Confinement

AGI confinement, or "AI boxing" (Drexler 1986; Vinge 1993; Chalmers 2010; Yampolskiy 2012; Armstrong, Sandberg, and Bostrom 2012), involves confining an AGI to a specific environment and limiting its access to the external world.

Yampolskiy (2012) makes an attempt to formalize the idea, drawing on previous computer security research on the so-called confinement problem (Lampson 1973). Yampolskiy defines the *AI confinement problem* (AICP) as the challenge of restricting an AGI to a confined environment from which it can't communicate without authorization.

27. Berglas (personal communication) has since changed his mind and no longer believes that it is possible to effectively restrict hardware or otherwise prevent AGI from being created.

A number of methods have been proposed for implementing AI confinement, many of which are extensively discussed in Armstrong, Sandberg, and Bostrom's (2012) paper.

Chalmers (2010) and Armstrong, Sandberg, and Bostrom (2012) mention numerous caveats and difficulties with AI-boxing approaches. A *truly* leakproof system that perfectly isolated the AGI from an outside environment would prevent us from even observing the AGI. If AGIs were given knowledge about human behavior or psychology, they could still launch social engineering attacks on us (Chalmers 2010; Armstrong, Sandberg, and Bostrom 2012). An AGI that was unaware of the existence of humans would be less likely to launch such attacks, but also much more limited in the kinds of tasks that it could be used for.

Even if the AGI remained confined, it could achieve enough influence among humans to prevent itself from being reset or otherwise modified (Good 1970). An AGI that people grew reliant on might also become impossible to reset or modify.

#### 4.1.1. Safe Questions

Yampolskiy (2012) proposes a hierarchy of security levels for communication with the AGI, ranging from completely unrestricted communication to no communication whatsoever. He suggests that the right level for communicating with AGIs is an intermediate one, based on asking "safe questions." These are defined as questions for which a human could find an answer without help from a superintelligence. For instance, a researcher who was considering two possible cures for cancer could ask the AGI which one to pursue first. Yampolskiy argues that the AGI is effectively not changing the final outcome, as the researcher would in principle be capable of finding the better treatment himself. The AGI, by recommending one of the approaches as the choice to pursue first, is merely speeding up the process. Armstrong, Sandberg, and Bostrom (2012) also briefly discuss such approaches.

#### 4.1.2. Virtual Worlds

While the confinement protocol supposes an AGI that can perceive our world directly, Chalmers (2010) uses the term "leakproof singularity" to refer to the project of creating isolated virtual worlds which would disallow their inhabitants from influencing the external world. An AGI placed in such a world could be monitored by humans to ensure that it was safe to release into our world. However, Chalmers notes that "for an AI system to be useful or interesting to us at all, it must have some effects on us. At a minimum, we must be able to observe it. And the moment we observe a virtual environment, some information leaks out from that environment into our environment and affects us."

Hibbard (2012b) proposes a decision support system for evaluating AGI designs in simulated environments that model our real environment, and attempts to show that the AGI can be built to be safe.

Armstrong, Sandberg, and Bostrom (2012) discuss the notion that we could supplement this approach with several layers of "sub-worlds": virtual worlds stacked within each other. In such a scenario, even if an AGI somehow manages to break out from one virtual reality, it is still restricted by another and cannot directly influence the world. Ultimately, the authors conclude that this approach isn't very promising.

#### 4.1.3. Resetting the AGI

For as long as an AGI remains confined or cooperative, its memory could be altered from the outside. Turney (1991) mentions a proposal in which an AGI's memory could be wiped after each job, thus limiting its potential. Armstrong, Sandberg, and Bostrom (2012) also mention the possibility of resetting an AGI after each run, though they note that there are many ways for the AGI to get around it, and that it is only useful for certain kinds of AGI systems.

Good (1970) warns that a superhuman AGI may be smart enough to find ways to prevent itself from being successfully reset or shut down, including methods of social manipulation. Warwick (1998) makes a similar point.

#### 4.1.4. Checks and Balances

Armstrong, Sandberg, and Bostrom (2012) suggest a number of checks and balances that could be used to catch the early stages of an AGI that was behaving dangerously or trying to deceive us. Test questions with known answers could be mixed in with questions we genuinely didn't know the answer to, in order to establish an AGI's trustworthiness. "Honey-pots," tempting resources with great value to the AGI, could also be left in physical or mental locations that it was forbidden from accessing. Finally, several different kinds of AGIs could be presented with the same problem, with their answers then being compared. Inconsistent answers would suggest a possible attempt at deception.

#### 4.1.5. "AI Confinement" Proposals: Our View

Despite their limited ability to deal with AGIs more intelligent than us, AI-boxing techniques seem to have value as a first line of defense, and it may be worthwhile to invest in developing off-the-shelf software packages for AI confinement that are easy and cheap to use. A research project that developed AGI unexpectedly might not have been motivated to make major investments in security, but the AGI might still be sufficiently limited in intelligence that confinement would work. Having a defense that is easy

to deploy will make it more likely that these kinds of projects will implement better precautions.

However, at the same time there is a risk that this will promote a false sense of security and make research teams think that they have carried out their duty to be cautious merely because they are running elementary confinement protocols. Although some confinement procedures can be implemented on top of an AGI that was not expressly designed for confinement, they are much less reliable than with an AGI that was built with confinement considerations in mind (Armstrong, Sandberg, and Bostrom 2012)—and even then, relying solely on confinement is a risky strategy. We are therefore somewhat cautious in our recommendation to develop confinement techniques.

### 4.2. AGI Enforcement

One problem with AI confinement proposals is that humans are tasked with guarding machines that may be far more intelligent than themselves (Good 1970). One proposed solution for this problem is to give the task of watching AGIs to other AGIs.

Armstrong (2007) proposes that the trustworthiness of a superintelligent system could be monitored via a chain of less powerful systems, all the way down to humans. Although humans couldn't verify and understand the workings of a superintelligence, they could verify and understand an AGI just slightly above their own level, which could in turn verify and understand an AGI somewhat above its own level, and so on.

Chaining multiple levels of AI systems with progressively greater capacity seems to be replacing the problem of building a safe AI with a multisystem, and possibly more difficult, version of the same problem. Armstrong himself admits that there are several problems with the proposal. There could be numerous issues along the line, such as a break in the chain of communication or an inability of a system to accurately assess the mind of another (smarter) system. There is also the problem of creating a trusted bottom for the chain in the first place, which is not necessarily any easier than creating a trustworthy superintelligence.

Hall (2007a) writes that there will be a great variety of AGIs, with those that were designed to be moral or aligned with human interests keeping the nonsafe ones in check. Goertzel and Pitt (2012) also propose that we build a community of mutually policing AGI systems of roughly equal levels of intelligence. If an AGI started to "go off the rails," the other AGIs could stop it. This might not prevent a single AGI from undergoing an intelligence explosion, but a community of AGIs might be in a better position to detect and stop it than humans would.

Having AGIs police each other is only useful if the group of AGIs actually has goals and values that are compatible with human goals and values. To this end, the appropriate internal constraints are needed.

The proposal of a society of mutually policing AGIs would avoid the problem of trying to control a more intelligent mind. If a global network of mildly superintelligent AGIs could be instituted in such a manner, it might detect and prevent any nascent takeoff. However, by itself such an approach is not enough to ensure safety—it helps guard against individual AGIs "going off the rails," but it does not help in a scenario where the programming of *most* AGIs is flawed and leads to nonsafe behavior. It thus needs to be combined with the appropriate internal constraints.

A complication is that a hard takeoff is a *relative* term—an event that happens too fast for any outside observer to stop. Even if the AGI network were a hundred times more intelligent than a network composed of only humans, there might still be a more sophisticated AGI that could overcome the network.

#### 4.2.1. "AGI Enforcement" Proposals: Our View

AGI enforcement proposals are in many respects similar to social integration proposals (section 3.2), in that they depend on the AGIs being part of a society which is strong enough to stop any single AGI from misbehaving. The greatest challenge is then to make sure that most of the AGIs in the overall system are safe and do not unite against humans rather than against misbehaving AGIs. Also, there might not be a natural distinction between a distributed AGI and a collection of many different AGIs, and AGI design is in any case likely to make heavy use of earlier AI/AGI techniques. AGI enforcement proposals therefore seem like implementation variants of various internal constraint proposals (section 5), rather than independent proposals.

### 4.3. Simulation Argument Attack

An external constraint loosely based on Bostrom's (2003a) simulation argument was proposed independently by Nelson (2007) and Miller (2012). The attack is based on the idea of attempting to convince any AGIs that they might be running within a computer simulation, and that any simulations exhibiting nonsafe behavior may be turned off. The AGI might then constrain itself to safe behavior in order to minimize the risk of being turned off, even if it was not actually being simulated.

Relevant for Nelson's formulation is the question of the rational course of action when one is uncertain about whether or not one is being simulated. A special case of this question is known in philosophy as the "Dr. Evil problem" (Elga 2004; Weatherson 2005).

#### 4.3.1. "Simulation Argument Attack" Proposals: Our View

Although the simulation argument attack is an interesting idea, AGIs would only be vulnerable to it if they were convinced that there was a nontrivial chance of being in a

simulation that might get shut down, and if their decision-making process considered it worth constraining their actions due to this possibility. It seems likely that a great many kinds of AGIs would fail either of these two criteria due to being intentionally designed to ignore such attacks, (correctly or incorrectly) considering the probability of being in a simulation to be too low to care about, not caring about simulations, or for any number of other reasons. They would then ignore the threat entirely. This makes it seem like the approach would have limited value at most, though it could be useful against some specific kinds of AGIs.

## 5. Internal Constraints

In addition to external constraints, AGIs could be designed with internal motivations designed to ensure that they would take actions in a manner beneficial to humanity. Alternatively, AGIs could be built with internal constraints that make them easier to control via external means.

With regard to internal constraints, Yudkowsky distinguishes between *technical failure* and *philosophical failure*:

> Technical failure is when you try to build an AI and it doesn't work the way you think it does—you have failed to understand the true workings of your own code. Philosophical failure is trying to build the wrong thing, so that even if you succeeded you would still fail to help anyone or benefit humanity. Needless to say, the two failures are not mutually exclusive. (Yudkowsky 2008a)

In practice, it is not always easy to distinguish between the two. Most of the discussion below focuses on the philosophical problems of various proposals, but some of the issues, such as whether or not a proposal is actually possible to implement, are technical.

### 5.1. Oracle AI

An *Oracle AI* is a hypothetical AGI that executes no actions other than answering questions. This might not be as safe as it sounds, however. Correctly defining "take no actions" might prove surprisingly tricky (Armstrong, Sandberg, and Bostrom 2012), and the oracle could give flawed advice even if it did correctly restrict its actions.

Some possible examples of flawed advice: As extra resources are useful for the fulfillment of nearly all goals (Omohundro 2007, 2008), the oracle may seek to obtain more resources—such as computing power—in order to answer questions more accurately. Its answers might then be biased toward furthering this goal, even if this temporarily reduces the accuracy of its answers, if it believes this to increase the accuracy of its answers in the long run. Another example is that if the oracle had the goal of answering

as many questions as possible as fast as possible, it could attempt to manipulate humans into asking it questions that were maximally simple and easy to answer.

Holden Karnofsky has suggested that an Oracle AI could be safe if it was "just a calculator," a system which only computed things that were asked of it, taking no goal-directed actions of its own. Such a "Tool-Oracle AI" would keep humans as the ultimate decision makers. Furthermore, the first team to create a Tool-Oracle AI could use it to become powerful enough to prevent the creation of other AGIs (Karnofsky and Tallinn 2011; Karnofsky 2012).

An example of a Tool-Oracle AI approach might be Omohundro's (2012) proposal of "Safe-AI Scaffolding": creating highly constrained AGI systems which act within limited, predetermined parameters. These could be used to develop formal verification methods and solve problems related to the design of more intelligent, but still safe, AGI systems.

### 5.1.1. Oracles Are Likely to Be Released

As with a boxed AGI, there are many factors that would tempt the owners of an Oracle AI to transform it to an autonomously acting agent. Such an AGI would be far more effective in furthering its goals, but also far more dangerous.

Current narrow-AI technology includes high-frequency trading (HFT) algorithms, which make trading decisions within fractions of a second, far too fast to keep humans in the loop. HFT seeks to make a very short-term profit, but even traders looking for a longer-term investment benefit from being faster than their competitors. Market prices are also very effective at incorporating various sources of knowledge (Hanson 2000). As a consequence, a trading algorithm's performance might be improved both by making it faster and by making it more capable of integrating various sources of knowledge. Most advances toward general AGI will likely be quickly taken advantage of in the financial markets, with little opportunity for a human to vet all the decisions. Oracle AIs are unlikely to remain as pure oracles for long.

Similarly, Wallach and Allen (2012) discuss the topic of autonomous robotic weaponry and note that the US military is seeking to eventually transition to a state where the human operators of robot weapons are "on the loop" rather than "in the loop." In other words, whereas a human was previously required to explicitly give the order before a robot was allowed to initiate possibly lethal activity, in the future humans are meant to merely supervise the robot's actions and interfere if something goes wrong.

Human Rights Watch (2012) reports on a number of military systems which are becoming increasingly autonomous, with the human oversight for automatic weapons defense systems—designed to detect and shoot down incoming missiles and rockets—already being limited to accepting or overriding the computer's plan of action in a matter

of seconds. Although these systems are better described as automatic, carrying out pre-programmed sequences of actions in a structured environment, than autonomous, they are a good demonstration of a situation where rapid decisions are needed and the extent of human oversight is limited. A number of militaries are considering the future use of more autonomous weapons.

In general, any broad domain involving high stakes, adversarial decision making, and a need to act rapidly is likely to become increasingly dominated by autonomous systems. The extent to which the systems will need general intelligence will depend on the domain, but domains such as corporate management, fraud detection, and warfare could plausibly make use of all the intelligence they can get. If one's opponents in the domain are also using increasingly autonomous AI/AGI, there will be an arms race where one might have little choice but to give increasing amounts of control to AI/AGI systems.

Miller (2012) also points out that if a person was close to death, due to natural causes, being on the losing side of a war, or any other reason, they might turn even a potentially dangerous AGI system free. This would be a rational course of action as long as they primarily valued their own survival and thought that even a small chance of the AGI saving their life was better than a near-certain death.

Some AGI designers might also choose to create less constrained and more free-acting AGIs for aesthetic or moral reasons, preferring advanced minds to have more freedom.

### 5.1.2. Oracles Will Become Authorities

Even if humans were technically kept in the loop, they might not have the time, opportunity, motivation, intelligence, or confidence to verify the advice given by an Oracle AI. This may be a danger even with narrower AI systems. Friedman and Kahn (1992) discuss APACHE, an expert system that provides medical advice to doctors. They write that as the medical community puts more and more trust into APACHE, it may become common practice to act automatically on APACHE's recommendations, and it may become increasingly difficult to challenge the "authority" of the recommendations. Eventually, APACHE may in effect begin to dictate clinical decisions.

Likewise, Bostrom and Yudkowsky (forthcoming) point out that modern bureaucrats often follow established procedures to the letter, rather than exercising their own judgment and allowing themselves to be blamed for any mistakes that follow. Dutifully following all the recommendations of an AGI system would be an even better way of avoiding blame.

Wallach and Allen (2012) note the existence of robots which attempt to automatically detect the locations of hostile snipers and to point them out to soldiers. To the extent

that these soldiers have come to trust the robots, they could be seen as carrying out the robots' orders. Eventually, equipping the robot with its own weapons would merely dispense with the formality of needing to have a human to pull the trigger.

Thus, even AGI systems that function purely to provide advice will need to be explicitly designed to be safe in the sense of not providing advice that would go against human values (Wallach and Allen 2009). Yudkowsky (2012) further notes that an Oracle AI might choose a plan that is beyond human comprehension, in which case there's still a need to design it as explicitly safe and conforming to human values.

#### 5.1.3. "Oracle AI" Proposals: Our View

Much like with external constraints, it seems like Oracle AIs could be a useful stepping stone on the path toward safe, freely acting AGIs. However, because any Oracle AI can be relatively easily turned into a free-acting AGI and because many people will have an incentive to do so, Oracle AIs are not by themselves a solution to AGI risk, even if they are safer than free-acting AGIs when kept as pure oracles.

### 5.2. Top-Down Safe AGI

AGIs built to take autonomous actions will need to be designed with safe motivations. Wallach and Allen divide approaches for ensuring safe behavior into "top-down" and "bottom-up" approaches (Allen, Varner, and Zinser 2000; Allen, Smit, and Wallach 2005; Wallach, Allen, and Smit 2008; Wallach and Allen 2009; Wallach 2010). They define "top-down" approaches as ones that take a specified ethical theory and attempt to build a system capable of implementing that theory (Wallach and Allen 2009).

Wallach and Allen (2000; 2005; 2008; 2009; 2010) have expressed skepticism about the feasibility of both pure top-down and bottom-up approaches, arguing for a hybrid approach.[28] With regard to top-down approaches, which attempt to derive an internal architecture from a given ethical theory, Wallach (2010) finds three problems:

1. "Limitations already recognized by moral philosophers: For example, in a utilitarian calculation, how can consequences be calculated when information is limited and the effects of actions cascade in never-ending interactions? Which consequences should be factored into the maximization of utility? Is there a stopping procedure?" (Wallach 2010)

2. The "frame problem": Each model of moral reasoning suffers from some version of the frame problem, the challenge of discerning relevant from irrelevant information without having to consider all of it, as all information could be relevant in principle

28. For a definition of "bottom-up" approaches, see section 5.3.

(Pylyshyn 1987; Dennett 1987). The requirements for psychological knowledge, knowledge of actions' effects in the world, and estimation of the sufficiency of the initial information all contribute to computational load.

3. "The need for background information. What mechanisms will the system require in order to acquire the information it needs to make its calculations? How does one ensure that this information is up to date in real time?" (Wallach 2010)

To some extent, these problems may be special cases of the fact that we do not yet have AGI with good general learning capabilities: creating an AGI would also require solving the frame problem, for instance. These problems might therefore not all be as challenging as they seem at first, presuming that we manage to develop AGI in the first place.

#### 5.2.1. Three Laws

Probably the most widely known proposal for machine ethics is Isaac Asimov's (1942) Three Laws of Robotics:

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.

Asimov and other writers later expanded the list to include a number of additional laws, including the Zeroth Law:

> A robot may not harm humanity, or through inaction allow humanity to come to harm.

Although the Three Laws are widely known and have inspired numerous imitations, several of Asimov's own stories were written to illustrate the fact that the laws contained numerous problems. They have also drawn heavy critique from others (Clarke 1993, 1994; Weld and Etzioni 1994; Pynadath and Tambe 2002; Gordon-Spears 2003; McCauley 2007; Weng, Chen, and Sun 2008; Wallach and Allen 2009; Murphy and Woods 2009; S. L. Anderson 2011) and are not considered a viable approach for safe AI. Among their chief shortcomings is the fact that they are too ambiguous to implement and, if defined with complete accuracy, contradict each other in many situations.

### 5.2.2. Categorical Imperative

The best-known universal ethical axiom is Kant's categorical imperative. Many authors have discussed using the categorical imperative as the foundation of AGI morality (Allen, Varner, and Zinser 2000; Stahl 2002; Powers 2006; Bugaj and Goertzel 2007; Wallach and Allen 2009; Beavers 2009, 2012). All of these authors conclude that Kantian ethics is a problematic goal for AGI, though Powers (2006) still remains hopeful about its prospects.

### 5.2.3. Principle of Voluntary Joyous Growth

Goertzel (2004a, 2004b) considers a number of possible axioms before settling on what he calls the "Principle of Voluntary Joyous Growth," defined as "Maximize happiness, growth and choice." He starts by considering the axiom "maximize happiness," but then finds this to be problematic and adds "growth," which he defines as "increase in the amount and complexity of patterns in the universe." Finally he adds "choice" in order to allow sentient beings to "choose their own destiny."

### 5.2.4. Utilitarianism

Classic utilitarianism is an ethical theory stating that people should take actions that lead to the greatest amount of happiness and the smallest amount of suffering. The prospects for AGIs implementing a utilitarian moral theory have been discussed by several authors (Good 1982; Gips 1995; Allen, Varner, and Zinser 2000; Anderson, Anderson, and Armen 2005c; Cloos 2005; Anderson, Anderson, and Armen 2006; Grau 2006; Wallach and Allen 2009; Muehlhauser and Helm 2012). The consensus is that pure classical utilitarianism is problematic and does not capture all human values. For example, a purely utilitarian AGI could reprogram the brains of humans so that they did nothing but experience the maximal amount of pleasure all the time, and that prospect seems unsatisfactory to many.[29]

### 5.2.5. Value Learning

Freeman (2009) describes a decision-making algorithm which observes people's behavior, infers their preferences in the form of a utility function, and then attempts to carry out actions which fulfill everyone's preferences. Similarly, Dewey (2011) discusses *value learners*, AGIs which are provided a probability distribution over possible utility

29. Note that utilitarianism is not the same thing as having a utility function. Utilitarianism is a specific kind of ethical system, while utility functions are general-purpose mechanisms for choosing between actions and can in principle be used to implement very different kinds of ethical systems, such as egoism and possibly even rights-based theories and virtue ethics (Peterson, Pisoni, and Miyamoto 2010).

functions that humans may have. Value learners then attempt to find the utility functions with the best match for human preferences. Hibbard (2012a) builds on Dewey's work to offer a similar proposal.

One problem with conceptualizing human desires as utility functions is that human desires change over time (van Gelder 1995) and also violate the axioms of utility theory required to construct a coherent utility function (Tversky and Kahneman 1981). While it is possible to treat inconsistent choices as random deviations from an underlying "true" utility function that is then learned (Nielsen and Jensen 2004), this does not seem to properly describe human preferences. Rather, human decision making and preferences seem to be driven by multiple competing systems, only some of which resemble utility functions (Dayan 2011). Even if a true utility function could be constructed, it does not take into account the fact that we have second-order preferences, or desires about our desires: a drug addict may desire a drug, but also desire that he not desire it (Frankfurt 1971). Similarly, we often wish that we had stronger desires toward behaviors which we consider good but cannot make ourselves engage in. Taking second-order preferences into account leads to what philosophers call "ideal preference" theories of value (Brandt 1979; Railton 1986; Lewis 1989; Sobel 1994; Zimmerman 2003; Tanyi 2006; Smith 2009).

Taking this into account, it has been argued that we should aim to build AGIs which act according to humanity's *extrapolated* values (Yudkowsky 2004; Tarleton 2010; Muehlhauser and Helm 2012). Yudkowsky proposes attempting to discover the "Coherent Extrapolated Volition" (CEV) of humanity, which he defines as

> our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted. (Yudkowsky 2004)

CEV remains vaguely defined and has been criticized by several authors (Hibbard 2005a; Goertzel 2010a; Waser 2011; Goertzel and Pitt 2012; Miller 2012). Tarleton (2010) suggests that CEV has five desirable properties, and that many different kinds of algorithms could possess these features:

**Meta-algorithm:** Most of the AGI's goals will be obtained at runtime from human minds, rather than explicitly programmed in before runtime.

**Factually correct beliefs:** The AGI will attempt to obtain correct answers to various factual questions, in order to modify preferences or desires that are based upon false factual beliefs.

**Singleton:** Only one superintelligent AGI is to be constructed, and it is to take control of the world with whatever goal function is decided upon.

**Reflection:** Individual or group preferences are reflected upon and revised.

**Preference aggregation:** The set of preferences of a whole group are to be combined somehow.

At least two CEV variants have been proposed: Coherent Aggregated Volition (Goertzel 2010a) and Coherent Blended Volition (Goertzel and Pitt 2012). Goertzel and Pitt (2012) describe the Coherent Blended Volition of a diverse group as analogous to a "conceptual blend," incorporating the most essential elements of the group into a harmonious whole. For this to be broadly acceptable, the different parties must agree that enough essential elements of their own views have been included.

Christiano (2012) attempts to sketch out a formalization of a value extrapolation approach called "indirect normativity." It proposes a technique that would allow an AI to approximate the kinds of values a group of humans would settle on if they could spend an unbounded amount of time and resources considering the problem.

Other authors have begun preliminary work on simpler value learning systems, designed to automatically learn moral principles. Anderson, Anderson, and Armen (2005a, 2005b, 2006) have built systems based around various moral duties and principles. As lists of duties do not in and of themselves specify what to do when they conflict, the systems let human experts judge how each conflict should be resolved, and then attempt to learn general rules from the judgments. As put forth, however, this approach would have little ability to infer ethical rules which did not fit the framework of proposed duties. Improved computational models of ethical reasoning (Ashley and McLaren 1995; McLaren 2003, 2006; Anderson, Anderson, and Armen 2005a, 2005b, 2006), perhaps incorporating work from neuroscience and moral psychology (Shulman, Jonsson, and Tarleton 2009b; Muehlhauser and Helm 2012; Bello and Bringsjord 2012), could help address this. Potapov and Rodionov (2012) propose an approach by which an AGI could gradually learn the values of other agents as its understanding of the world improved.

A value extrapolation process seems difficult to specify exactly, as it requires building an AGI with programming that formally and rigorously defines human values. Even if it manages to avoid the first issue in Wallach's (2010) list (section 5.2), top-down value extrapolation may fall victim to other issues, such as computational tractability. One interpretation of CEV would seem to require modeling not only the values of everyone on Earth, but also the evolution of those values as the people in question interacted with each other, became more intelligent and more like their ideal selves, chose which of their values they wanted to preserve, etc. Even if the AGI could eventually obtain the enormous amount of computing power required to run this model, its behavior would

need to be safe from the beginning, or it could end up doing vast damage to humanity before understanding what it was doing wrong.

Goertzel and Pitt's (2012) hybrid approach, in which AGIs cooperate with humans in order to discover the values humans wish to see implemented, seems more likely to avoid the issue of computational tractability. However, it will fail to work in a hard takeoff situation where AGIs take control before being taught the correct human values. Another issue with Coherent Blended Volition is that schemes which require absolute consensus are unworkable with large groups, as anyone whose situation would be worsened by a change of events could block the consensus. A general issue with value extrapolation approaches is that there may be several valid ways of defining a value extrapolation process, with no objective grounds for choosing one rather than another.

Goertzel (2010a) notes that in formal reasoning systems a set of initially inconsistent beliefs which the system attempts to resolve might arrive at something very different than the initial belief set, even if there existed a consistent belief set that was closer to the original set. He suggests that something similar might happen when attempting to make human values consistent, though whether this would be a bad thing is unclear.

#### 5.2.6. "Top-Down Safe AGI" Proposals: Our View

Of the various top-down proposals, value learning proposals seem to be the only ones which properly take into account the complexity of value thesis (section 2.2), as they attempt to specifically take into account considerations such as "Would humanity have endorsed this course of action if it had known the consequences?" Although there are many open questions concerning the computational tractability as well as the general feasibility of such approaches, they seem like some of the most important ones to work on.

### 5.3. Bottom-Up and Hybrid Safe AGI

Wallach (2010) defines bottom-up approaches as ones that favor evolving or simulating the mechanisms that give rise to our moral decisions. Another alternative is hybrid approaches, combining parts of both top-down and bottom-up approaches.

A problem with pure bottom-up approaches is that techniques such as artificial evolution or merely rewarding the AGI for the right behavior may cause it to behave correctly in tests, but would not guarantee that it would behave safely in any other situation. Even if an AGI *seems* to have adopted human values, the actual processes driving its behavior may be very different from the processes that would be driving the actions of a human who behaved similarly. It might then behave very unexpectedly in situations which are different enough (Yudkowsky 2008a, 2011; Ring and Orseau 2011; Hibbard 2012d).

Armstrong, Sandberg, and Bostrom discuss various problems related to such approaches and offer examples of concepts which seem straightforward to humans but are not as simple as they may seem on the surface. One of their examples relates to the concept of time:

> If the [AGI] had the reasonable-sounding moral premise that "painlessly killing a human being, who is going to die in a micro-second anyway, in order to gain some other good, is not a crime," we would not want it to be able to redefine millennia as seconds. (Armstrong, Sandberg, and Bostrom 2012)

All humans have an intuitive understanding of time and no experience with beings who could arbitrarily redefine their own clocks and might not share the same concept of time. Such differences in the conceptual grounding of an AGI's values and of human values might not become apparent until too late.

### 5.3.1. Evolutionary Invariants

Human morality is to a large extent shaped by evolution (de Waal et al. 2006; Joyce 2001), and evolutionary approaches attempt to replicate this process with AGIs.

Hall (2007a, 2011) argues that self-improving AGIs are likely to exist in competition with many other kinds of self-improving AGIs. Properties that give AGIs a significant disadvantage might then be strongly selected against and disappear. We could attempt to identify *evolutionary invariants*, or evolutionarily stable strategies, which would both survive in a competitive environment and cause an AGI to treat humans well.

Hall (2011) lists self-interest, long planning horizons, knowledge, an understanding of evolutionary ethics, and guaranteed honesty as invariants that are likely to make an AGI more moral as well as to persist even under intense competition. He suggests that, although self-interest may sound like a bad thing in an AGI, non-self-interested creatures are difficult to punish. Thus, enlightened self-interest might be a good thing for an AGI to possess, as it will provide an outside community with both a stick and a carrot to control it with.

Similarly, Waser (2008) suggests that minds which are intelligent enough will, due to game-theoretical and other considerations, become altruistic and cooperative. Waser (2011) proposed the principle of Rational Universal Benevolence (RUB), the idea that the moral course of action is cooperation while letting everyone freely pursue their own goals. Waser proposes that, instead of making human-friendly behavior an AGI's only goal, the AGI would be allowed to have and form its own goals. However, its goals and actions would be subject to the constraint that they should respect the principle of RUB and not force others into a life those others would disagree with.

Kornai (forthcoming) cites Gewirth's (1978) work on the principle of generic consistency, which holds that respecting others' rights to freedom and well-being is a logically necessary conclusion for any rational agents. Kornai suggests that if the principle is correct, then AGIs would respect humanity's rights to freedom and well-being, and that AGIs which failed to respect the principle would be outcompeted by ones which did.

Something similar was also proposed by Gips (1995) and Versenyi (1974), who advocates the creation of "wise robots" that would recognize the extent to which their own well-being depended on cooperation with humans, and would act accordingly.

Although these approaches expect AGI either to evolve altruism or to find it the most rational approach, true altruism or even pure tit-for-tat (Axelrod 1987) isn't actually the best strategy in evolutionary terms. Rather, a better strategy is *Machiavellian* tit-for-tat: cultivating an appearance of altruism and cooperation when it benefits oneself, and acting selfishly when one can get away with it. Humans seem strongly disposed toward such behavior (Haidt 2006).

Another problem is that tit-for-tat as a good strategy assumes that both players are equally powerful and both have the same options at their disposal. If the AGI became far more powerful than most humans, it might no longer be in its interests to treat humans favorably (Fox and Shulman 2010). This hypothesis can be tested by looking at human behavior: if exploiting the weak is an evolutionarily useful strategy, then humans should engage in it when given the opportunity. Humans who feel powerful do indeed devalue the worth of the less powerful and view them as objects of manipulation (Kipnis 1972). They also tend to ignore social norms (van Kleef et al. 2011) and to experience less distress and compassion toward the suffering of others (van Kleef et al. 2008).

Thus, even if an AGI cooperated with other similarly powerful AGIs, the group of AGIs might still decide to collectively exploit humans. Similarly, even though there might be pressure for AGIs to make themselves more transparent and easily inspected by others, this only persists for as long as the AGI needs others more than the others need the AGI.

### 5.3.2. Evolved Morality

Another proposal is to create AGIs via algorithmic evolution, selecting in each generation the AGIs which are not only the most intelligent, but also the most moral. These ideas are discussed to some extent by Wallach and Allen (2009).

### 5.3.3. Reinforcement Learning

In machine learning, *reinforcement learning* is a model in which an agent takes various actions and is differentially rewarded for the actions, after which it learns to perform the actions with the greatest expected reward. In psychology, it refers to agents being

rewarded for certain actions and thus learning behaviors which they have a hard time breaking, even if some other kind of behavior is more beneficial for them later on.

Applied to AGI, the machine learning sense of reinforcement involves teaching an AGI to behave in a safe manner by rewarding it for ethical choices, and letting it learn for itself the underlying rules of what constitutes ethical behavior. In an early example of this kind of proposal, McCulloch (1956) described an "ethical machine" that could infer the rules of chess by playing the game, and suggested that it could also learn ethical behavior this way.

Hibbard (2001, 2005a) suggested using reinforcement learning to give AGIs positive emotions toward humans. Early AGIs would be taught to recognize happiness and unhappiness in humans, and the results of this learning would be hard-wired as emotional values in more advanced AGIs. This training process would then be continued—for example, by letting the AGIs predict how human happiness would be affected by various actions and using those predictions as emotional values.

A reinforcement learner is supplied with a reward signal, and it always has the explicit goal of maximizing the sum of this reward, any way it can. In order for this goal to align with human values, humans must engineer the environment so that the reinforcement learner is prevented from receiving rewards if human goals are not fulfilled (Dewey 2011). A reinforcement-learning AGI only remains safe for as long as humans are capable of enforcing this limitation, and will become unpredictable if it becomes capable of overcoming it. Hibbard (2012d) has retracted his earlier reinforcement learning–based proposals, as they would allow the AGI to maximize its reinforcement by modifying humans to be maximally happy, even against their will (Dewey 2011).

### 5.3.4. Human-like AGI

Another kind of proposal involves building AGIs that can learn human values by virtue of being similar to humans.

Connectionist systems, based on artificial neural nets, are capable of learning patterns from data without being told what the patterns are. As some connectionist models have learned to classify problems in a manner similar to humans (McLeod, Plunkett, and Rolls 1998; Plaut 2003; Thomas and McClelland 2008), it has been proposed that connectionist AGI might learn moral principles that are too complex for humans to specify as explicit rules.[30] This idea has been explored by Guarini (2006) and Wallach and Allen (2009).

30. But it should be noted that there are also promising nonconnectionist approaches for modeling human classification behavior: see, e.g., Tenenbaum, Griffiths, and Kemp (2006).

One specific proposal that draws upon connectionism is to make AGIs act according to virtue ethics (Allen, Varner, and Zinser 2000; Wallach, Allen, and Smit 2008; Wallach and Allen 2009; Wallach 2010). These authors note that previous writers discussing virtuous behavior have emphasized the importance of learning moral virtues through habit and practice. As it is impossible to exhaustively define a virtue, virtue ethics has traditionally required each individual to learn the right behaviors through "bottom-up processes of discovery or learning" (Wallach and Allen 2009). Models that mimicked the human learning process well enough could then potentially learn the same behaviors as humans do.

Another kind of human-inspired proposal is the suggestion that something like Stan Franklin's LIDA architecture (Franklin and Patterson 2006; Ramamurthy et al. 2006; Snaider, Mccall, and Franklin 2011), or some other approach based on Bernard Baars's (2002, 2005) "global workspace" theory, might enable moral reasoning. In the LIDA architecture, incoming information is monitored by specialized *attention codelets*, each of which searches the input for specific features. In particular, *moral* codelets might look for morally relevant factors and ally themselves with other codelets to promote their concerns to the level of conscious attention. Ultimately, some coalitions will win enough support to accomplish a specific kind of decision (Wallach and Allen 2009; Wallach, Franklin, and Allen 2010; Wallach, Allen, and Franklin 2011).

Goertzel and Pitt (2012) consider human memory systems (episodic, sensorimotor, declarative, procedural, attentional, and intentional) and ways by which human morality might be formed via their interaction. They briefly discuss the way that the OpenCog AGI system (Hart and Goertzel 2008; Goertzel 2012a) implements similar memory systems and how those systems could enable it to learn morality. Similarly, Goertzel and Bugaj (2008) discuss the stages of moral development in humans and suggest ways by which they could be replicated in AGI systems, using the specific example of Novamente, a proprietary version of OpenCog.

Waser (2009) also proposes building an AGI by studying the results of evolution and creating an implementation as close to the human model as possible.

Human-inspired AGI architectures would intuitively seem the most capable of learning human values, though what would be human-like enough remains an open question. It is possible that even a relatively minor variation from the norm could cause an AGI to adopt values that most humans would consider undesirable. Getting the details right might require an extensive understanding of the human brain.

There are also humans who have drastically different ethics than the vast majority of humanity and argue for the desirability of outcomes such as the extinction of mankind (Benatar 2006; Dietrich 2007). There remains the possibility that even AGIs which reasoned about ethics in a completely human-like manner would reach such conclusions.

Humans also have a long track record of abusing power, or of undergoing major behavioral changes due to relatively small injuries—the "safe *Homo sapiens*" problem also remains unsolved. On the other hand, it seems plausible that human-like AGIs could be explicitly engineered to avoid such problems.

The easier that an AGI is to modify, the more powerful it might become (Sotala 2012), and very close recreations of the human brain may turn out to be difficult to extensively modify and upgrade. Human-inspired safe AGIs might then end up outcompeted by AGIs which were easier to modify, and which might or might not be safe.

Even if human-inspired architectures could be easily modified, the messiness of human cognitive architecture means that it might be difficult to ensure that their values remain beneficial during modification. For instance, in LIDA-like architectures, beneficial behavior will depend on the correct coalitions of morality codelets winning each time. If the system undergoes drastic changes, this can be very difficult if not impossible to ensure.

Most AGI builders will attempt to create a mind that displays considerable advantages over ordinary humans. Some such advantages might be best achieved by employing a very nonhuman architecture (Moravec 1992), so there will be reasons to build AGIs that are not as human-like. These could also end up outcompeting the human-like AGIs.

#### 5.3.5. "Bottom-Up and Hybrid Safe AGI" Proposals: Our View

We are generally very skeptical about pure bottom-up methods, as they only allow a very crude degree of control over an AGI's goals, giving it a motivational system which can only be relied on to align with human values in the very specific environments that the AGI has been tested in. Evolutionary invariants seem incapable of preserving complexity of value, and they might not even be capable of preserving human survival. Reinforcement learning, on the other hand, depends on the AGI being incapable of modifying the environment against the will of its human controllers. Therefore, none of these three approaches seems workable.

Human-like AGI might have some promise, depending on exactly how fragile human values were. If the AGI reasoning process could be made sufficiently human-like, there is the possibility that the AGI could remain relatively safe, though less safe than a well-executed value extrapolation–based AGI.

### 5.4. AGI Nanny

A more general proposal, which could be achieved by either top-down, bottom-up, or hybrid methods, is the proposal of an "AGI Nanny" (Goertzel 2012b; Goertzel and Pitt 2012). This is an AGI that is somewhat more intelligent than humans and is designed to monitor Earth for various dangers, including more advanced AGI.

The AGI Nanny would be connected to powerful surveillance systems and would control a massive contingent of robots. It would help abolish problems such as disease, involuntary death, and poverty, while preventing the development of technologies that could threaten humanity. The AGI Nanny would be designed not to rule humanity on a permanent basis, but to give us some breathing room and time to design more advanced AGIs. After some predetermined amount of time, it would cede control of the world to a more intelligent AGI.

Goertzel and Pitt (2012) briefly discuss some of the problems inherent in the AGI Nanny proposal. The AGI would have to come to power in an ethical way, and might behave unpredictably despite our best efforts. It might also be easier to create a dramatically self-improving AGI than to create a more constrained AGI Nanny.

#### 5.4.1. "AGI Nanny" Proposals: Our View

Upon asserting control, the AGI Nanny would need to have precisely specified goals, so that it would stop other AGIs from taking control but would also not harm human interests. It is not clear to what extent defining these goals would be easier than defining the goals of a more free-acting AGI (Muehlhauser and Salamon 2012). Overall, the AGI Nanny seems to have promise, but it's unclear whether it can be made to work.

### 5.5. Formal Verification

Formal verification methods prove specific properties about various algorithms. If the complexity and fragility of value theses hold, it follows that safe AGI requires the ability to verify that proposed changes to the AGI will not alter its goals or values. If even a mild drift from an AGI's original goals might lead to catastrophic consequences, then utmost care should be given to ensuring that the goals will not change inadvertently. This is particularly the case if there are no external feedback mechanisms which would correct the drift. Before modifying itself, an AGI could attempt to formally prove that the changes would not alter its existing goals, and would therefore keep them intact even during extended self-modification (Yudkowsky 2008a). Such proofs could be required before self-modification was allowed to occur, and the system could also be required to prove that this verify-before-modification property itself would always be preserved during self-modification.

Formal verification is also an essential part of Omohundro's (2012) Safe-AI Scaffolding strategy, as noted in section 5.6.4.

Spears (2006) combines machine learning and formal verification methods to ensure that AIs remain within the bounds of prespecified constraints after having learned new behaviors. She attempts to identify "safe" machine learning operators, which are guaranteed to preserve the system's constraints.

One AGI system built entirely around the concept of formal verification is the Gödel machine (Schmidhuber 2009; Steunebrink and Schmidhuber 2011). It consists of a *solver*, which attempts to achieve the goals set for the machine, and a *searcher*, which has access to a set of axioms that completely describe the machine. The searcher may completely rewrite any part of the machine, provided that it can produce a formal proof showing that such a rewrite will further the system's goals.

Goertzel (2010b) proposes GOLEM (Goal-Oriented LEarning Meta-Architecture), a meta-architecture that can be wrapped around a variety of different AGI systems. GOLEM will only implement changes that are predicted to be more effective at achieving the original goal of the system. Goertzel argues that GOLEM is likely to be both self-improving and steadfast: either it pursues the same goals it had at the start, or it stops acting altogether.

Unfortunately, formalizing the AGI's goals in a manner that will allow formal verification methods to be used is a challenging task. Within cryptography, many communications protocols have been proven secure, only for successful attacks to be later developed against their various implementations. While the formal proofs were correct, they contained assumptions which did not accurately capture the way the protocols worked in practice (Degabriele, Paterson, and Watson 2011). Proven theorems are only as good as their assumptions, so formal verification requires good models of the AGI hardware and software.

### 5.5.1. "Formal Verification" Proposals: Our View

Compared to the relatively simple domain of cryptographic security, verifying things such as "Does this kind of a change to the AGI's code preserve its goal of respecting human values?" seems like a much more open-ended and difficult task, one which might even prove impossible. Regardless, it is the only way of achieving high confidence that a system is safe, so it should at least be attempted.

## 5.6. Motivational Weaknesses

Finally, there is a category of internal constraints that, while not making an AGI's *values* safer, make it easier to control AGI via external constraints.

### 5.6.1. High Discount Rates

AGI systems could be given a high discount rate, making them value short-term goals and gains far more than long-term goals and gains (Shulman 2010a; Armstrong, Sandberg, and Bostrom 2012). This would inhibit the AGI's long-term planning, making it more predictable. However, an AGI can also reach long-term goals through a series of short-term goals (Armstrong, Sandberg, and Bostrom 2012).

### 5.6.2. Easily Satiable Goals

Shulman (2010a) proposes designing AGIs in such a way that their goals are easy to satisfy. For example, an AGI could receive a near-maximum reward for simply continuing to receive an external reward signal, which could be cut if humans suspected misbehavior. The AGI would then prefer to cooperate with humans rather than trying to attack them, even if it was very sure of its chances of success.[31] Likewise, if the AGI could receive a maximal reward with a relatively small fraction of humanity's available resources, it would have little to gain from seizing more resources.

An extreme form of this kind of a deal is Orseau and Ring's (2011) "Simpleton Gambit," in which an AGI is promised everything that it would ever want, on the condition that it turn itself into a harmless simpleton. Orseau and Ring consider several hypothetical AGI designs, many of which seem likely to accept the gambit, given certain assumptions.

In a related paper, Ring and Orseau (2011) consider the consequences of an AGI being able to modify itself to receive the maximum possible reward. They show that certain types of AGIs will then come to only care about their own survival. Hypothetically, humans could promise not to threaten such AGIs in exchange for them agreeing to be subject to AI-boxing procedures. For this to work, the system would have to believe that humans will take care of its survival against external threats better than it could itself. Hibbard (2012a, 2012c) discusses the kinds of AGIs that would avoid the behaviors described by Ring and Orseau (2011; 2011).

### 5.6.3. Calculated Indifference

Another proposal is to make an AGI indifferent to a specific event. For instance, the AGI could be made indifferent to the detonation of explosives attached to its hardware, which might enable humans to have better control over it (Armstrong 2010; Armstrong, Sandberg, and Bostrom 2012).

### 5.6.4. Programmed Restrictions

Goertzel and Pitt (2012) suggest we ought to ensure that an AGI does not self-improve too fast, because AGIs will be harder to control as they become more and more cognitively superior to humans. To limit the rate of self-improvement in AGIs, perhaps AGIs could be programmed to extensively consult humans and other AGI systems

31. On the other hand, this might incentivize the AGI to deceive its controllers into believing it was behaving properly, and also to actively hide any information which it even suspected might be interpreted as misbehavior.

while improving themselves, in order to ensure that no unwanted modifications would be implemented.

Omohundro (2012) discusses a number of programmed restrictions in the form of constraints on what the AGI is allowed to do, with formal proofs being used to ensure that an AGI will not violate its safety constraints. Such limited AGI systems would be used to design more sophisticated AGIs.

Programmed restrictions are problematic, as the AGI might treat these merely as problems to solve in the process of meeting its goals, and attempt to overcome them (Omohundro 2008). Making an AGI not want to quickly self-improve might not solve the problem by itself. If the AGI ends up with a second-order desire to rid itself of such a disinclination, the stronger desire will eventually prevail (Suber 2002). Even if the AGI wanted to maintain its disinclination toward rapid self-improvement, it might still try to circumvent the goal in some other way, such as by creating a copy of itself which did not have that disinclination (Omohundro 2008). Regardless, such limits could help control less sophisticated AGIs.

#### 5.6.5. Legal Machine Language

Weng, Chen, and Sun (2008, 2009) propose a "legal machine language" which could be used to formally specify actions which the AGI is allowed or disallowed to do. Governments could then enact laws written in legal machine language, allowing them to be programmed into robots.

#### 5.6.6. "Motivational Weaknesses" Proposals: Our View

Overall, motivational weaknesses seem comparable to external constraints: possibly useful and worth studying, but not something to rely on exclusively, particularly in the case of superintelligent AGIs. As with external constraints and Oracle AIs, an arms race situation might provide a considerable incentive to loosen or remove such constraints.

| Responses to Catastrophic AGI Risk | | | |
|---|---|---|---|
| **Societal Proposals** | | | |
| **Do nothing** | | AGI is distant | |
| | | Little risk, no action needed | |
| | | Let them kill us | |
| **Integrate to society** | | Legal and economic controls | |
| | | Foster positive values | |
| **Regulate research** | | Review boards | |
| | | Encourage safety research | |
| | | Differential technological progress | |
| | | International mass surveillance | |
| **Enhance human capabilities** | | | |
| **Relinquish technology** | | Outlaw AGI | |
| | | Restrict hardware | |
| **AGI design proposals** | | | |
| **External constraints** | | **Internal constraints** | |
| **AGI confinement** | Safe questions | **Oracle AI** | |
| | Virtual worlds | **Top-down approaches** | Three laws |
| | Resetting the AGI | | Categorical imperative |
| | Checks and balances | | Principle of Voluntary Joyous Growth |
| | | | Utilitarianism |
| | | | Value learning |
| **AGI enforcement** | | **Bottom-up and hybrid approaches** | Evolutionary invariants |
| | | | Evolved morality |
| | | | Reinforcement learning |
| | | | Human-like AGI |
| **Simulation argument attack** | | **AGI Nanny** | |
| | | **Formal verification** | |
| | | **Motivational weaknesses** | High discount rates |
| | | | Easily satiable goals |
| | | | Calculated indifference |
| | | | Programmed restrictions |
| | | | Legal Machine Language |

## 6. Conclusion

We began this paper by noting that a number of researchers are predicting AGI in the next twenty to one hundred years. One must not put excess trust in this time frame: as Armstrong and Sotala (2012) show, experts have been terrible at predicting AGI. Muehlhauser and Salamon (2012) consider a number of methods other than expert opinion that could be used for predicting AGI, but find that they too provide suggestive evidence at best.

It would be a mistake, however, to leap from "AGI is very hard to predict" to "AGI must be very far away." Our brains are known to think about uncertain, abstract ideas like AGI in "far mode," which also makes it feel like AGI must be temporally distant (Trope and Liberman 2010; Muehlhauser 2012), but something being *uncertain* is not strong evidence that it is *far away*. When we are highly ignorant about something, we should widen our error bars in both directions. Thus, we shouldn't be highly confident that AGI will arrive this century, and we shouldn't be highly confident that it *won't*.

Next, we explained why AGIs may be an existential risk. A trend toward automatization would give AGIs increased influence in society, and there might be a discontinuity in which they gained power rapidly. This could be a disaster for humanity if AGIs don't share our values, and in fact it looks difficult to cause them to share our values because human values are complex and fragile, and therefore problematic to specify.

The recommendations given for dealing with the problem can be divided into proposals for societal action (section 3), external constraints (section 4), and internal constraints (section 5). Many proposals seem to suffer from serious problems, or seem to be of limited effectiveness. Others seem to have enough promise to be worth exploring. We will conclude by reviewing the proposals which we feel are worthy of further study.

As a brief summary of our views, in the medium term, we think that the proposals of AGI confinement (section 4.1), Oracle AI (section 5.1), and motivational weaknesses (section 5.6) would have promise in helping create safer AGIs. These proposals share in common the fact that, although they could help a cautious team of researchers create an AGI, they are not solutions to the problem of AGI risk, as they do not prevent others from creating unsafe AGIs, nor are they sufficient in guaranteeing the safety of sufficiently intelligent AGIs. Regulation (section 3.3) as well as human capability enhancement (section 3.4) could also help to somewhat reduce AGI risk. In the long run, we will need the ability to guarantee the safety of freely acting AGIs. For this goal, value learning (section 5.2.5) would seem like the most reliable approach if it could be made to work, with human-like architecture (section 5.3.4) a possible alternative which seems less reliable but possibly easier to build. Formal verification (section 5.5) seems

like a very important tool in helping to ensure the safety of our AGIs, regardless of the exact approach that we choose.

Of the societal proposals, we are supportive of the calls to regulate AGI development, but we admit there are many practical hurdles which might make this infeasible. The economic and military potential of AGI, and the difficulty of verifying regulations and arms treaties restricting it, could lead to unstoppable arms races.

We find ourselves in general agreement with the authors who advocate funding additional research into safe AGI as the primary solution. Such research will also help establish the kinds of constraints which would make it possible to successfully carry out integration proposals.

Uploading approaches, in which human minds are made to run on computers and then augmented, might buy us some time to develop safe AGI. However, it is unclear whether they can be developed before AGI, and large-scale uploading could create strong evolutionary trends which seem dangerous in and of themselves. As AGIs seem likely to eventually outpace uploads, uploading by itself is probably not a sufficient solution. What uploading could do is to reduce the initial advantages that AGIs enjoy over (partially uploaded) humanity, so that other responses to AGI risk can be deployed more effectively.

External constraints are likely to be useful in controlling AGI systems of limited intelligence, and could possibly help us develop more intelligent AGIs while maintaining their safety. If inexpensive external constraints were readily available, this could encourage even research teams skeptical about safety issues to implement them. Yet it does not seem safe to rely on these constraints once we are dealing with a superhuman intelligence, and we cannot trust everyone to be responsible enough to contain their AGI systems, especially given the economic pressures to "release" AGIs. For such an approach to be a solution for AGI risk in general, it would have to be adopted by all successful AGI projects, at least until safe AGIs were developed. Much the same is true of attempting to design Oracle AIs. In the short term, such efforts may be reinforced by research into motivational weaknesses, internal constraints that make AGIs easier to control via external means.

In the long term, the internal constraints that show the most promise are value extrapolation approaches and human-like architectures. Value extrapolation attempts to learn human values and interpret them as we would wish them to be interpreted. These approaches have the advantage of potentially maximizing the preservation of human values, and the disadvantage that such approaches may prove intractable, or impossible to properly formalize. Human-like architectures seem easier to construct, as we can simply copy mechanisms that are used within the human brain, but it seems hard to build such an exact match as to reliably replicate human values. Slavish reproductions

of the human psyche also seem likely to be outcompeted by less human, more efficient architectures.

Both approaches would benefit from better formal verification methods, so that AGIs which were editing and improving themselves could verify that the modifications did not threaten to remove the AGIs' motivation to follow their original goals. Studies which aim to uncover the roots of human morals and preferences also seem like candidates for research that would benefit the development of safe AGI (Shulman, Jonsson, and Tarleton 2009b; Muehlhauser and Helm 2012; Bello and Bringsjord 2012), as do studies into computational models of ethical reasoning (McLaren 2006).

We reiterate that when we talk about "human values," we are not making the claim that human values would be static, nor that *current* human values would be ideal. Nor do we wish to imply that the values of other sentient beings would be unimportant. Rather, we are seeking to guarantee the implementation of some very basic values, such as the avoidance of unnecessary suffering, the preservation of humanity, and the prohibition of forced brain reprogramming. We believe the vast majority of humans would agree with these values and be sad to see them lost.

## Acknowledgments

Special thanks to Luke Muehlhauser for extensive assistance throughout the writing process. We would also like to thank Abram Demski, Alexei Turchin, Alexey Potapov, Anders Sandberg, Andras Kornai, Anthony Berglas, Aron Vallinder, Ben Goertzel, Ben Noble, Ben Sterrett, Brian Rabkin, Bill Hibbard, Carl Shulman, Dana Scott, Daniel Dewey, David Pearce, Evelyn Mitchell, Evgenij Thorstensen, Frank White, gwern branwen, Harri Valpola, Jaan Tallinn, Jacob Steinhardt, James Babcock, James Miller, Joshua Fox, Louie Helm, Mark Gubrud, Mark Waser, Michael Anissimov, Michael Vassar, Miles Brundage, Moshe Looks, Randal Koene, Robin Hanson, Risto Saarelma, Steve Omohundro, Steven Kaas, Stuart Armstrong, Tim Freeman, Ted Goertzel, Toni Heinonen, Tony Barrett, Vincent Müller, Vladimir Nesov, Wei Dai, and several users of LessWrong.com for their helpful comments.

## References

Adams, Sam S., Itamar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, John Storrs Hall, et al. 2012. "Mapping the Landscape of Human-Level Artificial General Intelligence." *AI Magazine* 33 (1): 25–42.

Agar, Nicholas. 2011. "Ray Kurzweil and Uploading: Just Say No!" *Journal of Evolution and Technology* 22 (1): 23–36. http://jetpress.org/v22/agar.pdf.

Agliata, Daniel, and Stacey Tantleff-Dunn. 2004. "The Impact of Media Exposure on Males' Body Image." *Journal of Social and Clinical Psychology* 23 (1): 7–22. doi:10.1521/jscp.23.1.7.26988.

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, and Wendell Wallach. 2012. "Moral Machines: Contradiction in Terms or Abdication of Human Responsibility." In Lin, Abney, and Bekey 2012, 55–68.

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

Amdahl, Gene M. 1967. "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities." In *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference—AFIPS '67 (Spring),* 483–485. New York: ACM Press. doi:10.1145/1465482.1465560.

Anderson, Michael, and Susan Leigh Anderson, eds. 2011. *Machine Ethics.* New York: Cambridge University Press.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen, eds. 2005a. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Technical Report, FS-05-06. AAAI Press, Menlo Park, CA. http://www.aaai.org/Library/Symposia/Fall/fs05-06.

———. 2005b. "MedEthEx: Toward a Medical Ethics Advisor." In *Caring Machines: AI in Eldercare: Papers from the 2005 AAAI Fall Symposium,* edited by Timothy Bickmore, 9–16. Technical Report, FS-05-02. AAAI Press, Menlo Park, CA. http://aaaipress.org/Papers/Symposia/Fall/2005/FS-05-02/FS05-02-002.pdf.

———. 2005c. "Towards Machine Ethics: Implementing Two Action-Based Ethical Theories." In Anderson, Anderson, and Armen 2005a, 1–7.

———. 2006. "An Approach to Computing Ethics." *IEEE Intelligent Systems* 21 (4): 56–63. doi:10.1109/MIS.2006.64.

Anderson, Monica. 2010. "Problem Solved: Unfriendly AI." *H+ Magazine,* December 15. http://hplusmagazine.com/2010/12/15/problem-solved-unfriendly-ai/.

Anderson, Susan Leigh. 2011. "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics." In Anderson and Anderson 2011, 285–296.

Annas, George J., Lori B. Andrews, and Rosario M. Isasi. 2002. "Protecting the Endangered Human: Toward an International Treaty Prohibiting Cloning and Inheritable Alterations." *American Journal of Law & Medicine* 28 (2–3): 151–178.

Anthony, Dick, and Thomas Robbins. 2004. "Conversion and 'Brainwashing' in New Religious Movements." In *The Oxford Handbook of New Religious Movements,* 1st ed., edited by James R. Lewis, 243–297. New York: Oxford University Press. doi:10.1093/oxfordhb/9780195369649.003.0012.

Arkin, Ronald C. 2009. *Governing Lethal Behavior in Autonomous Robots.* Boca Raton, FL: CRC Press.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Accessed December 31, 2012. http://www.neweuropeancentury.org/GodAI.pdf.

———. 2010. *Utility Indifference.* Technical Report, 2010-1. Oxford: Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/reports/2010-1.pdf.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. "Thinking Inside the Box: Controlling and Using an Oracle AI." *Minds and Machines* 22 (4): 299–324. doi:10.1007/s11023-012-9282-2.

Armstrong, Stuart, and Kaj Sotala. 2012. "How We're Predicting AI — or Failing To." In *Beyond AI: Artificial Dreams,* edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia. Accessed February 2, 2013. http://www.kky.zcu.cz/en/publications/1/JanRomportl_2012_BeyondAIArtificial.pdf.

Asaro, Peter M. 2007. "Robots and Responsibility from a Legal Perspective." In *Proceedings of the IEEE Conference on Robotics and Automation, Workshop on Roboethics.* Rome, April 14. http://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf.

Ashley, Kevin D., and Bruce M. McLaren. 1995. "Reasoning with Reasons in Case-Based Comparisons." In *Proceedings of the First International Conference on Case-Based Reasoning Research and Development,* edited by Manuela M. Veloso and Agnar Aamodt, 133–144. Berlin: Springer. http://www.cs.cmu.edu/~bmclaren/pubs/AshleyMcLaren-ReasoningWithReasons-ICCBR95.pdf.

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

Axelrod, Robert. 1987. "The Evolution of Strategies in the Iterated Prisoner's Dilemma." In *Genetic Algorithms and Simulated Annealing*, edited by Lawrence Davis, 32–41. Los Altos, CA: Morgan Kaufmann.

Baars, Bernard J. 2002. "The Conscious Access Hypothesis: Origins and Recent Evidence." *Trends in Cognitive Sciences* 6 (1): 47–52. doi:10.1016/S1364-6613(00)01819-2.

———. 2005. "Global Workspace Theory of Consciousness: Toward a Cognitive Neuroscience of Human Experience." In *The Boundaries of Consciousness: Neurobiology and Neuropathology*, edited by Steven Laureys, 45–53. Progress in Brain Research 150. Boston: Elsevier.

Bach, Joscha, Ben Goertzel, and Matthew Iklé, eds. 2012. *Artificial General Intelligence: 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings.* Lecture Notes in Artificial Intelligence 7716. New York: Springer. doi:10.1007/978-3-642-35506-6.

Bamford, Sim. 2012. "A Framework for Approaches to Transfer of a Mind's Substrate." *International Journal of Machine Consciousness* 4 (1): 23–34. doi:10.1142/S1793843012400021.

Baum, Seth D., Ben Goertzel, and Ted G. Goertzel. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185–195. doi:10.1016/j.techfore.2010.09.006.

Beavers, Anthony F. 2009. "Between Angels and Animals: The Question of Robot Ethics; or, Is Kantian Moral Agency Desirable?" Paper presented at the Annual Meeting of the Association for Practical and Professional Ethics, Cincinnati, OH, March.

———. 2012. "Moral Machines and the Threat of Ethical Nihilism." In Lin, Abney, and Bekey 2012, 333–344.

Bello, Paul, and Selmer Bringsjord. 2012. "On How to Build a Moral Machine." *Topoi*. doi:10.1007/s11245-012-9129-8.

Benatar, David. 2006. *Better Never to Have Been: The Harm of Coming into Existence*. New York: Oxford University Press.

Berglas, Anthony. 2012. "Artificial Intelligence Will Kill Our Grandchildren (Singularity)." Unpublished manuscript, draft 9, January. Accessed December 31, 2012. http://berglas.org/Articles/AIKillGrandchildren/AIKillGrandchildren.html.

Blackmore, Susan. 2012. "She Won't Be Me." *Journal of Consciousness Studies* 19 (1–2): 16–19. http://www.susanblackmore.co.uk/Articles/JCS2012.htm.

Bostrom, Nick. 1998. "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.

———. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. http://www.jetpress.org/volume9/risks.html.

———. 2003a. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243–255. doi:10.1111/1467-9213.00309.

———. 2003b. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy, 339–371. Vol. 2. Death and Anti-Death. Palo Alto, CA: Ria University Press.

———. 2007. "Technological Revolutions: Ethics and Policy in the Dark." In *Nanoscale: Issues and Perspectives for the Nano Century*, edited by Nigel M. de S. Cameron and M. Ellen Mitchell, 129–152. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9780470165874.ch10.

———. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Müller. Special issue, *Minds and Machines* 22 (2): 71–85. doi:10.1007/s11023-012-9281-3.

Bostrom, Nick, and Milan M. Ćirković, eds. 2008a. *Global Catastrophic Risks.* New York: Oxford University Press.

———. 2008b. "Introduction." In Bostrom and Ćirković 2008a, 1–30.

Bostrom, Nick, and Eliezer Yudkowsky. Forthcoming. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.

Brain, Marshall. 2003. "Robotic Nation." Accessed December 31, 2012. http://marshallbrain.com/robotic-nation.htm.

Brandt, Richard B. 1979. *A Theory of the Good and the Right.* New York: Oxford University Press.

Branwen, Gwern. 2012. "Slowing Moore's Law: Why You Might Want to and How You Would Do It." gwern.net. December 11. Accessed December 31, 2012. http://www.gwern.net/Slowing%20Moore's%20Law.

Brin, David. 1998. *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* Reading, MA: Perseus Books.

Bringsjord, Selmer, and Alexander Bringsjord. 2012. "Belief in the Singularity is Fideistic." In Eden, Søraker, Moor, and Steinhart 2012.

Brooks, Rodney A. 2008. "I, Rodney Brooks, Am a Robot." *IEEE Spectrum* 45 (6): 68–71. doi:10.1109/MSPEC.2008.4531466.

Brynjolfsson, Erik, and Andrew McAfee. 2011. *Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy.* Lexington, MA: Digital Frontier. Kindle edition.

Bryson, Joanna J. 2010. "Robots Should be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues,* edited by Yorick Wilks, 107–126. Philadelphia, PA: John Benjamins.

Bryson, Joanna, and Phil Kime. 1998. "Just Another Artifact: Ethics and the Empirical Experience of AI." Paper presented at the Fifteenth Internation Congress on Cybernetics, Namur, Belgium. http://www.cs.bath.ac.uk/~jjb/web/aiethics98.html.

Bugaj, Stephan Vladimir, and Ben Goertzel. 2007. "Five Ethical Imperatives and Their Implications for Human-AGI Interaction." *Dynamical Psychology.* http://goertzel.org/dynapsyc/2007/Five_Ethical_Imperatives_svbedit.htm.

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.

Cade, C. Maxwell. 1966. *Other Worlds Than Ours.* 1st ed. London: Museum.

Calandrino, Joseph A., William Clarkson, and Edward W. Felten. 2011. "Bubble Trouble: Off-Line De-Anonymization of Bubble Forms." In *Proceedings of the 20th USENIX Security Symposium*, 267–280. San Francisco, CA: USENIX. http://www.usenix.org/events/sec11/tech/full_papers/Calandrino.pdf.

Cassimatis, Nicholas, Erik T. Mueller, and Patrick Henry Winston. 2006. "Achieving Human-Level Intelligence Through Integrated Systems and Research: Introduction to This Special Issue." In "Achieving Human-Level Intelligence Through Integrated Systems and Research." *AI Magazine* 27 (2): 12–14. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1876/1774.

Casti, John L. 2012. *X-Events: The Collapse of Everything*. New York: William Morrow.

Cattell, Rick, and Alice Parker. 2012. *Challenges for Brain Emulation: Why is Building a Brain so Difficult?* Synaptic Link, February 5. http://synapticlink.org/Brain%20Emulation%20Challenges.pdf.

CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission). 2010. *Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC. http://www.sec.gov/news/studies/2010/marketevents-report.pdf.

Chalmers, David John. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Philosophy of Mind Series. New York: Oxford University Press.

———. 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.

Christiano, Paul F. 2012. "'Indirect Normativity' Write-up." *Ordinary Ideas* (blog), April 21. http://ordinaryideas.wordpress.com/2012/04/21/indirect-normativity-write-up/.

Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. 1st ed. 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.

Cloos, Christopher. 2005. "The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism." In Anderson, Anderson, and Armen 2005a, 38–45.

Dahm, Werner J. A. 2010. *Technology Horizons: A Vision for Air Force Science & Technology During 2010-2030*. AF/ST-TR-10-01-PR. Washington, DC: USAF. http://www.au.af.mil/au/awc/awcgate/af/tech_horizons_vol-1_may2010.pdf.

Daley, William. 2011. "Mitigating Potential Hazards to Humans from the Development of Intelligent Machines." *Synthese* 2:44–50. http://www.synesisjournal.com/vol2_g/2011_2_44-50_Daley.pdf.

Davis, Ernest. 2012. "The Singularity and the State of the Art in Artificial Intelligence." Working Paper, New York, May 9. Accessed July 22, 2013. http://www.cs.nyu.edu/~davise/papers/singularity.pdf.

Dayan, Peter. 2011. "Models of Value and Choice." In *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms,* edited by Raymond J. Dolan and Tali Sharot, 33–52. Waltham, MA: Academic Press.

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.

De Waal, Frans, Robert Wright, Christine M. Korsgaard, Philip Kitcher, and Peter Singer. 2006. *Primates and Philosophers: How Morality Evolved.* 1st ed. Edited by Stephen Macedo and Josiah Ober. Princeton, NJ: Princeton University Press.

Degabriele, Jean Paul, Kenny Paterson, and Gaven Watson. 2011. "Provable Security in the Real World." *IEEE Security & Privacy Magazine* 9 (3): 33–41. doi:10.1109/MSP.2010.200.

Dennett, Daniel C. 1987. "Cognitive Wheels: The Frame Problem of AI." In Pylyshyn 1987, 41–64.

———. 2012. "The Mystery of David Chalmers." *Journal of Consciousness Studies* 19 (1–2): 86–95. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00005.

Deutsch, David. 2011. *The Beginning of Infinity: Explanations that Transform the World.* 1st ed. New York: Viking.

Dewey, Daniel. 2011. "Learning What to Value." In Schmidhuber, Thórisson, and Looks 2011, 309–314.

Dietrich, Eric. 2007. "After The Humans Are Gone." *Philosophy Now,* May–June. http://philosophynow.org/issues/61/After_The_Humans_Are_Gone.

Docherty, Bonnie, and Steve Goose. 2012. *Losing Humanity: The Case Against Killer Robots.* Cambridge, MA: Human Rights Watch and the International Human Rights Clinic, November 19. http://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf.

Douglas, Thomas. 2008. "Moral Enhancement." *Journal of Applied Philosophy* 25 (3): 228–245. doi:10.1111/j.1468-5930.2008.00412.x.

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

Eckersley, Peter, and Anders Sandberg. Forthcoming. "Is Brain Emulation Dangerous?"

Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. 2012. *Singularity Hypotheses: A Scientific and Philosophical Assessment.* The Frontiers Collection. Berlin: Springer.

Eisen, Michael. 2011. "Amazon's $23,698,655.93 Book about Flies." *It is NOT Junk* (blog), April 22. http://www.michaeleisen.org/blog/?p=358.

Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69 (2): 383–396. doi:10.1111/j.1933-1592.2004.tb00400.x.

Felten, Edward W., and Michael A. Schneider. 2000. "Timing Attacks on Web Privacy." In *Proceedings of the 7th ACM Conference on Computer and Communications Security - CCS '00,* 25–32. New York: ACM Press.

Ferguson, Melissa J., Ran Hassin, and John A. Bargh. 2007. "Implicit Motivation: Past, Present, and Future." In *Handbook of Motivation Science,* edited by James Y. Shah and Wendi L. Gardner, 150–166. New York: Guilford.

Fox, Joshua, and Carl Shulman. 2010. "Superintelligence Does Not Imply Benevolence." In Mainzer 2010.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20. doi:10.2307/2024717.

Franklin, Stan, and F. G. Patterson Jr. 2006. "The LIDA Architecture: Adding New Modes of Learning to an Intelligent, Autonomous, Software Agent." In *IDPT-2006 Proceedings*. San Diego, CA: Society for Design & Process Science. http://ccrg.cs.memphis.edu/assets/papers/zo-1010-lida-060403.pdf.

Freeman, Tim. 2008. "Comparative Advantage Doesn't Ensure Survival." Unpublished manuscript, November 23. Accessed December 31, 2012. http://www.fungible.com/comparative-advantage.html.

———. 2009. "Using Compassion and Respect to Motivate an Artificial Intelligence." Unpublished manuscript, March 8. Accessed December 31, 2012. http://fungible.com/respect/paper.html.

Friedman, Batya, and Peter H. Kahn. 1992. "Human Agency and Responsible Computing: Implications for Computer System Design." *Journal of Systems and Software* 17 (1): 7–14. doi:10.1016/0164-1212(92)90075-U.

Gewirth, Alan. 1978. *Reason and Morality*. Chicago: University of Chicago Press.

Gips, James. 1995. "Towards the Ethical Robot." In *Android Epistemology*, edited by Kenneth M. Ford, Clark N. Glymour, and Patrick J. Hayes, 243–252. Cambridge, MA: MIT Press.

Goertzel, Ben. 2002. "Thoughts on AI Morality." *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2002/AIMorality.htm.

———. 2004a. "Encouraging a Positive Transcension: Issues in Transhumanist Ethical Philosophy." *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2004/PositiveTranscension.htm.

———. 2004b. "Growth, Choice and Joy: Toward a Precise Definition of a Universal Ethical Principle." *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2004/GrowthChoiceJoy.htm.

———. 2006. "Apparent Limitations on the 'AI Friendliness' and Related Concepts Imposed By the Complexity of the World." Working Paper, September. Accessed December 31, 2012. http://www.goertzel.org/papers/LimitationsOnFriendliness.pdf.

———. 2010a. "Coherent Aggregated Volition: A Method for Deriving Goal System Content for Advanced, Beneficial AGIs." *The Multiverse According to Ben* (blog), March 12. http://multiverseaccordingtoben.blogspot.ca/2010/03/coherent-aggregated-volition-toward.html.

———. 2010b. "GOLEM: Toward an AGI Meta-Architecture Enabling Both Goal Preservation and Radical Self-Improvement." Unpublished manuscript, May 2. Accessed December 31, 2012. http://goertzel.org/GOLEM.pdf.

———. 2012a. "CogPrime: An Integrative Architecture for Embodied Artificial General Intelligence." OpenCog Foundation. October 2. Accessed December 31, 2012. http://wiki.opencog.org/w/CogPrime_Overview.

———. 2012b. "Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?" *Journal of Consciousness Studies* 19 (1–2): 96–111. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00006.

———. 2012c. "When Should Two Minds Be Considered Versions of One Another?" *International Journal of Machine Consciousness* 4 (1): 177–185. doi:10.1142/S1793843012400094.

Goertzel, Ben, and Stephan Vladimir Bugaj. 2008. "Stages of Ethical Development in Artificial General Intelligence Systems." In Wang, Goertzel, and Franklin 2008, 448–459.

Goertzel, Ben, and Joel Pitt. 2012. "Nine Ways to Bias Open-Source AGI Toward Friendliness." *Journal of Evolution and Technology* 22 (1): 116–131. http://jetpress.org/v22/goertzel-pitt.htm.

Golle, Philippe, and Kurt Partridge. 2009. "On the Anonymity of Home/Work Location Pairs." In *Pervasive Computing,* edited by Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. Brush, and Yoshito Tobe, 390–397. Lecture Notes in Computer Science 5538. Springer. doi:10.1007/978-3-642-01516-8_26.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

———. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1–4): 67–79. doi:10.1080/00207237008709398.

———. 1982. "Ethical Machines." In *Intelligent Systems: Practice and Perspective,* edited by J. E. Hayes, Donald Michie, and Y.-H. Pao, 555–560. Machine Intelligence 10. Chichester: Ellis Horwood.

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,* edited by 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.

Groesz, Lisa M., Michael P. Levine, and Sarah K. Murnen. 2001. "The Effect of Experimental Presentation of Thin Media Images on Body Satisfaction: A Meta-Analytic Review." *International Journal of Eating Disorders* 31 (1): 1–16. doi:10.1002/eat.10005.

Guarini, Marcello. 2006. "Particularism and the Classification and Reclassification of Moral Cases." *IEEE Intelligent Systems* 21 (4): 22–28. doi:10.1109/MIS.2006.76.

Gubrud, Mark Avrum. 1997. "Nanotechnology and International Security." Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, November 5–8. http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/.

Gunkel, David J. 2012. *The Machine Question: Critical Perspectives on AI, Robotics, and Ethics.* Cambridge, MA: MIT Press.

Guterl, Fred. 2012. *The Fate of the Species: Why the Human Race May Cause Its Own Extinction and How We Can Stop It.* 1st ed. New York: Bloomsbury.

Haidt, Jonathan. 2006. *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom.* 1st ed. New York: Basic Books.

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

———. 2007b. "Ethics for Artificial Intellects." In *Nanoethics: The Ethical and Social Implications of Nanotechnology,* edited by Fritz Allhoff, Patrick Lin, James Moor, John Weckert, and Mihail C. Roco, 339–352. Hoboken, NJ: John Wiley & Sons.

———. 2008. "Engineering Utopia." In Wang, Goertzel, and Franklin 2008, 460–467.

Hall, John Storrs. 2011. "Ethics for Self-Improving Machines." In Anderson and Anderson 2011, 512–523.

Hallevy, Gabriel. 2010. "The Criminal Liability of Artificial Intelligence Entities." Unpublished manuscript, February 15. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564096.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

———. 1998. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Accessed May 15, 2013. http://hanson.gmu.edu/aigrow.pdf.

———. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Unpublished manuscript, September. Last revised October 2007. http://hanson.gmu.edu/futarchy.pdf.

———. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45–50. doi:10.1109/MSPEC.2008.4531461.

———. 2009. "Prefer Law to Values." *Overcoming Bias* (blog), October 10. http://www.overcomingbias.com/2009/10/prefer-law-to-values.html.

———. 2012. "Meet the New Conflict, Same as the Old Conflict." *Journal of Consciousness Studies* 19 (1–2): 119–125. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00008.

Hare, Robert D., Danny Clark, Martin Grann, and David Thornton. 2000. "Psychopathy and the Predictive Validity of the PCL-R: An International Perspective." *Behavioral Sciences & the Law* 18 (5): 623–645. doi:10.1002/1099-0798(200010)18:5<623::AID-BSL409>3.0.CO;2-W.

Harris, Grant T., and Marnie E. Rice. 2006. "Treatment of Psychopathy: A Review of Empirical Findings." In *Handbook of Psychopathy,* edited by Christopher J. Patrick, 555–572. New York: Guilford.

Hart, David, and Ben Goertzel. 2008. "OpenCog: A Software Framework for Integrative Artificial General Intelligence." Unpublished manuscript. http://www.agiri.org/OpenCog_AGI-08.pdf.

Hauskeller, Michael. 2012. "My Brain, My Mind, and I: Some Philosophical Assumptions of Mind-Uploading." *International Journal of Machine Consciousness* 4 (1): 187–200. doi:10.1142/S1793843012400100.

Hayworth, Kenneth J. 2012. "Electron Imaging Technology for Whole Brain Neural Circuit Mapping." *International Journal of Machine Consciousness* 4 (1): 87–108. doi:10.1142/S1793843012500060.

Heylighen, Francis. 2007. "Accelerating Socio-Technological Evolution: From Ephemeralization and Stigmergy to the Global Brain." In *Globalization as Evolutionary Process: Modeling Global Change,* edited by George Modelski, Tessaleno Devezas, and William R. Thompson, 284–309. Rethinking Globalizations 10. New York: Routledge.

———. 2012. "Brain in a Vat Cannot Break Out." *Journal of Consciousness Studies* 19 (1–2): 126–142. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00009.

Hibbard, Bill. 2001. "Super-Intelligent Machines." *ACM SIGGRAPH Computer Graphics* 35 (1): 13–15. http://www.siggraph.org/publications/newsletter/issues/v35/v35n1.pdf.

———. 2005a. "Critique of the SIAI Collective Volition Theory." Unpublished manuscript, December. Accessed December 31, 2012. http://www.ssec.wisc.edu/~billh/g/SIAI_CV_critique.html.

———. 2005b. "The Ethics and Politics of Super-Intelligent Machines." Unpublished manuscript, July. Microsoft Word file, accessed December 31, 2012. https://sites.google.com/site/whibbard/g/SI_ethics_politics.doc.

———. 2008. "Open Source AI." In Wang, Goertzel, and Franklin 2008, 473–477.

———. 2012a. "Avoiding Unintended AI Behaviors." In Bach, Goertzel, and Iklé 2012, 107–116.

———. 2012b. "Decision Support for Safe AI Design." In Bach, Goertzel, and Iklé 2012, 117–125.

———. 2012c. "Model-Based Utility Functions." *Journal of Artificial General Intelligence* 3 (1): 1–24. doi:10.2478/v10229-011-0013-5.

———. 2012d. *The Error in My 2001 VisFiles Column,* September. Accessed December 31, 2012. http://www.ssec.wisc.edu/~billh/g/visfiles_error.html.

Hollerbach, John M., Matthew T. Mason, and Henrik I. Christensen. 2009. *A Roadmap for US Robotics: From Internet to Robotics.* Snobird, UT: Computing Community Consortium. http://www.us-robotics.us/reports/CCC%20Report.pdf.

Hopkins, Patrick D. 2012. "Why Uploading Will Not Work, or, the Ghosts Haunting Transhumanism." *International Journal of Machine Consciousness* 4 (1): 229–243. doi:10.1142/S1793843012400136.

Horvitz, Eric J., and Bart Selman. 2009. *Interim Report from the AAAI Presidential Panel on Long-Term AI Futures.* Palo Alto, CA: AAAI, August. http://www.aaai.org/Organization/Panel/panel-note.pdf.

Hughes, James. 2001. "Relinquishment or Regulation: Dealing with Apocalyptic Technological Threats." Hartford, CT, November 14.

Hutter, Marcus. 2012. "Can Intelligence Explode?" *Journal of Consciousness Studies* 19 (1–2): 143–166. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00010.

*IEEE Spectrum.* 2008. "Tech Luminaries Address Singularity": "The Singularity; Special Report." (June).

Jenkins, Anne. 2003. "Artificial Intelligence and the Real World." *Futures* 35 (7): 779–786. doi:10.1016/S0016-3287(03)00029-6.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired,* April. http://www.wired.com/wired/archive/8.04/joy.html.

Joyce, Richard. 2001. *The Evolution of Morality.* Cambridge Studies in Philosophy. New York: Cambridge University Press. doi:10.2277/0521808065.

Karnofsky, Holden. 2012. "Thoughts on the Singularity Institute (SI)." *Less Wrong* (blog), May 11. http://lesswrong.com/lw/cbs/thoughts_on_the_singularity_institute_si/.

Karnofsky, Holden, and Jaan Tallinn. 2011. "Karnofsky & Tallinn Dialog on SIAI Efficacy." Accessed December 31, 2012. http://xa.yimg.com/kq/groups/23070378/1331435883/name/Jaan+Tallinn+2011+05+-+revised.doc.

Kipnis, David. 1972. "Does Power Corrupt?" *Journal of Personality and Social Psychology* 24 (1): 33–41. doi:10.1037/h0033390.

Koene, Randal A. 2012a. "Embracing Competitive Balance: The Case for Substrate-Independent Minds and Whole Brain Emulation." In Eden, Søraker, Moor, and Steinhart 2012.

———. 2012b. "Experimental Research in Whole Brain Emulation: The Need for Innovative in Vivo Measurement Techniques." *International Journal of Machine Consciousness* 4 (1): 35–65. doi:10.1142/S1793843012400033.

Kornai, Andras. Forthcoming. "Bounding the Impact of AGI." Paper presented at the AGI Impacts conference 2012.

Kringelbach, Morten L., and Kent C. Berridge, eds. 2009. *Pleasures of the Brain.* Series in Affective Science. New York: Oxford University Press.

Kurzweil, Ray. 2001. "Response to Stephen Hawking." Kurzweil Accelerating Intelligence. September 5. Accessed December 31, 2012. http://www.kurzweilai.net/response-to-stephen-hawking.

———. 2002. "Locked in His Chinese Room: Response to John Searle." In *Are We Spiritual Machines?: Ray Kurzweil vs. the Critics of Strong AI,* edited by Jay W. Richards, 128–171. Seattle, WA: Discovery Institute. http://www.kurzweilai.net/chapter-6-locked-in-his-chinese-room-response-to-john-searle.

———. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Lampson, Butler W. 1973. "A Note on the Confinement Problem." *Communications of the ACM* 16 (10): 613–615. doi:10.1145/362375.362389.

Legg, Shane. 2009. "Funding Safe AGI." *Vetta Project* (blog), August 3. http://www.vetta.org/2009/08/funding-safe-agi/.

Legg, Shane, and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence." In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop 2006,* edited by Ben Goertzel and Pei Wang, 17–24. Frontiers in Artificial Intelligence and Applications 157. Amsterdam: IOS.

Lehman-Wilzig, Sam N. 1981. "Frankenstein Unbound: Towards a Legal Definition of Artificial Intelligence." *Futures* 13 (6): 442–457. doi:10.1016/0016-3287(81)90100-2.

Levy, David. 2009. "The Ethical Treatment of Artificially Conscious Robots." *International Journal of Social Robotics* 1 (3): 209–216. doi:10.1007/s12369-009-0022-6.

Lewis, David. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volumes* 63:113–137. http://www.jstor.org/stable/4106918.

Lin, Patrick, Keith Abney, and George A. Bekey, eds. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics.* Intelligent Robotics and Autonomous Agents. Cambridge, MA: MIT Press.

Loosemore, Richard, and Ben Goertzel. 2012. "Why an Intelligence Explosion is Probable." In Eden, Søraker, Moor, and Steinhart 2012.

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

Mann, Steve, Jason Nolan, and Barry Wellman. 2003. "Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments." *Surveillance & Society* 1 (3): 331–355. http://library.queensu.ca/ojs/index.php/surveillance-and-society/article/view/3344.

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.

McCulloch, W. S. 1956. "Toward Some Circuitry of Ethical Robots; or, An Observational Science of the Genesis of Social Evaluation in the Mind-like Behavior of Artifacts." *Acta Biotheoretica* 11 (3–4): 147–156. doi:10.1007/BF01557008.

McDermott, Drew. 2012. "Response to 'The Singularity' by David Chalmers." *Journal of Consciousness Studies* 19 (1–2): 167–172. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00011.

McGinnis, John O. 2010. "Accelerating AI." *Northwestern University Law Review* 104 (3): 1253–1270. http://www.law.northwestern.edu/lawreview/v104/n3/1253/LR104n3McGinnis.pdf.

McKibben, Bill. 2003. *Enough: Staying Human in an Engineered Age.* New York: Henry Holt.

McLaren, Bruce M. 2003. "Extensionally Defining Principles and Cases in Ethics: An AI Model." *Artificial Intelligence* 150 (1–2): 145–181. doi:10.1016/S0004-3702(03)00135-8.

———. 2006. "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions." *IEEE Intelligent Systems* 21 (4): 29–37. doi:10.1109/MIS.2006.67.

McLeod, Peter, Kim Plunkett, and Edmund T. Rolls. 1998. *Introduction to Connectionist Modelling of Cognitive Processes.* New York: Oxford University Press.

Meuer, Hans, Erich Strohmaier, Jack Dongarra, and Horst Simon. 2012. "Top500 List - November 2012." TOP500 Supercomputer Sites. November. Accessed December 31, 2012. http://www.top500.org/list/2012/11/.

Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World.* Dallas, TX: BenBella Books.

Minsky, Marvin, Push Singh, and Aaron Sloman. 2004. "The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence." *AI Magazine* 25 (2): 113–124. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1764.

Moore, David, Vern Paxson, Stefan Savage, Colleen Shannon, Stuart Staniford, and Nicholas Weaver. 2003. "Inside the Slammer Worm." *IEEE Security & Privacy Magazine* 1 (4): 33–39. doi:10.1109/MSECP.2003.1219056.

Moore, David, Colleen Shannon, and Jeffery Brown. 2002. "Code-Red: A Case Study on the Spread and Victims of an Internet Worm." In *Proceedings of the Second ACM SIGCOMM Workshop on Internet Measurment (IMW '02),* 273–284. New York: ACM Press. doi:10.1145/637201.637244.

Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence.* Cambridge, MA: Harvard University Press.

———. 1992. "Pigs in Cyberspace." Field Robotics Center. Accessed December 31, 2012. http://www.frc.ri.cmu.edu/~hpm/project.archive/general.articles/1992/CyberPigs.html.

———. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1. http://www.transhumanist.com/volume1/moravec.htm.

———. 1999. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.

Moskowitz, Gordon B., Peizhong Li, and Elizabeth R. Kirk. 2004. "The Implicit Volition Model: On the Preconscious Regulation of Temporarily Adopted Goals." *Advances in Experimental Social Psychology* 36:317–413. doi:10.1016/S0065-2601(04)36006-5.

Muehlhauser, Luke. 2012. "Overconfident Pessimism." *Less Wrong* (blog), November 24. http://lesswrong.com/lw/fmf/overconfident_pessimism/.

Muehlhauser, Luke, and Louie Helm. 2012. "The Singularity and Machine Ethics." In Eden, Søraker, Moor, and Steinhart 2012.

Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In Eden, Søraker, Moor, and Steinhart 2012.

Mueller, Dennis C. 2003. *Public Choice III.* 3rd ed. New York: Cambridge University Press.

Murphy, Robin, and David D. Woods. 2009. "Beyond Asimov: The Three Laws of Responsible Robotics." *IEEE Intelligent Systems* 24 (4): 14–20. doi:10.1109/MIS.2009.69.

Napier, William. 2008. "Hazards from Comets and Asteroids." In Bostrom and Ćirković 2008a, 222–237.

Narayanan, Arvind, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. "On the Feasibility of Internet-Scale Author Identification." In *2012 IEEE Symposium on Security and Privacy,* 300–314. San Francisco, CA: IEEE Computer Society. doi:10.1109/SP.2012.46.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets." In *2008 IEEE Symposium on Security and Privacy,* 111–125. Oakland, CA: IEEE Computer Society. doi:10.1109/SP.2008.33.

———. 2009a. "De-anonymizing Social Networks." In *30th IEEE Symposium on Security and Privacy,* 173–187. Berkeley, CA: IEEE Computer Society. doi:10.1109/SP.2009.22.

———. 2009b. "De-Anonymizing Social Networks: FAQ." May 8. Accessed December 31, 2012. http://www.cs.utexas.edu/~shmat/socialnetworks-faq.html.

Nelson, Rolf. 2007. "How to Deter a Rogue AI by Using Your First-mover Advantage." SL4. August 22. Accessed December 31, 2012. http://www.sl4.org/archive/0708/16600.html.

Nielsen, Thomas D., and Finn V. Jensen. 2004. "Learning a Decision Maker's Utility Function from (Possibly) Inconsistent Behavior." *Artificial Intelligence* 160 (1–2): 53–78. doi:10.1016/j.artint.2004.08.003.

Nordmann, Alfred. 2007. "If and Then: A Critique of Speculative NanoEthics." *NanoEthics* 1 (1): 31–46. doi:10.1007/s11569-007-0007-6.

———. 2008. "Singular Simplicity." *IEEE Spectrum,* June. http://spectrum.ieee.org/robotics/robotics-software/singular-simplicity.

Olson, Mancur. 1982. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities.* New Haven, CT: Yale University Press.

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, CA, September 8–9. http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/.

———. 2008. "The Basic AI Drives." In Wang, Goertzel, and Franklin 2008, 483–492.

———. 2012. "Rational Artificial Intelligence for the Greater Good." In Eden, Søraker, Moor, and Steinhart 2012.

Orseau, Laurent, and Mark Ring. 2011. "Self-Modification and Mortality in Artificial Agents." In Schmidhuber, Thórisson, and Looks 2011, 1–10.

Persson, Ingmar, and Julian Savulescu. 2008. "The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity." *Journal of Applied Philosophy* 25 (3): 162–177. doi:10.1111/j.1468-5930.2008.00410.x.

———. 2012. *Unfit for the Future.* Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199653645.001.0001.

Peterson, Nathaniel R., David B. Pisoni, and Richard T. Miyamoto. 2010. "Cochlear Implants and Spoken Language Processing Abilities: Review and Assessment of the Literature." *Restorative Neurology and Neuroscience* 28 (2): 237–250. doi:10.3233/RNN-2010-0535.

Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

Plaut, David C. 2003. "Connectionist Modeling of Language: Examples and Implications." In *Mind, Brain, and Language: Multidisciplinary Perspectives,* edited by Marie T. Banich and Molly Mack, 143–168. Mahwah, NJ: Lawrence Erlbaum.

Posner, Richard A. 2004. *Catastrophe: Risk and Response.* New York: Oxford University Press.

Potapov, Alexey, and Sergey Rodionov. 2012. "Universal Empathy and Ethical Bias for Artificial General Intelligence." Paper presented at the Fifth Conference on Artificial General Intelligence (AGI–12), Oxford, December 8–11. Accessed June 27, 2013. http://aideus.com/research/doc/preprints/04_paper4_AGIImpacts12.pdf.

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

———. 2011. "Incremental Machine Ethics." *IEEE Robotics & Automation Magazine* 18 (1): 51–58. doi:10.1109/MRA.2010.940152.

Pylyshyn, Zenon W., ed. 1987. *The Robot's Dilemma: The Frame Problem in Artificial Intelligence.* Norwood, NJ: Ablex.

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,* edited by John-Jules Ch. Meyer and Milind Tambe, 307–320. Berlin: Springer. doi:10.1007/3-540-45448-9_22.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5–31.

Rajab, Moheeb Abu, Jay Zarfoss, Fabian Monrose, and Andreas Terzis. 2007. "My Botnet is Bigger than Yours (Maybe, Better than Yours): Why Size Estimates Remain Challenging." In *Proceedings of 1st Workshop on Hot Topics in Understanding Botnets (HotBots '07).* Berkeley, CA: USENIX. http://static.usenix.org/event/hotbots07/tech/full_papers/rajab/rajab.pdf.

Ramamurthy, Uma, Bernard J. Baars, Sidney K. D'Mello, and Stan Franklin. 2006. "LIDA: A Working Model of Cognition." In *Proceedings of the Seventh International Conference on Cognitive Modeling,* edited by Danilo Fum, Fabio Del Missier, and Andrea Stocco, 244–249. Trieste, Italy: Edizioni Goliardiche. http://ccrg.cs.memphis.edu/assets/papers/ICCM06-UR.pdf.

Reynolds, Carson, and Alvaro Cassinelli, eds. 2009. *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings.* AP-CAP 2009. http://kant.k2.t.u-tokyo.ac.jp/ap-cap09/proceedings.pdf.

Ring, Mark, and Laurent Orseau. 2011. "Delusion, Survival, and Intelligent Agents." In Schmidhuber, Thórisson, and Looks 2011, 11–20.

Salekin, Randall T. 2010. "Treatment of Child and Adolescent Psychopathy: Focusing on Change." In *Handbook of Child and Adolescent Psychopathy,* edited by Randall T. Salekin and Donald R. Lynam, 343–373. New York: Guilford.

Sandberg, Anders. 2001. "Friendly Superintelligence." Accessed December 31, 2012. http://www.aleph.se/Nada/Extro5/Friendly%20Superintelligence.htm.

———. 2010. "An Overview of Models of Technological Singularity." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8. http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf.

———. 2012. "Models of a Singularity." In Eden, Søraker, Moor, and Steinhart 2012.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf.

———. 2011. *Machine Intelligence Survey.* Technical Report, 2011-1. Future of Humanity Institute, University of Oxford. www.fhi.ox.ac.uk/reports/2011-1.pdf.

Schmidhuber, Jürgen. 2009. "Ultimate Cognition *à la* Gödel." *Cognitive Computation* 1 (2): 177–193. doi:10.1007/s12559-009-9014-y.

Schmidhuber, Jürgen, Kristinn R. Thórisson, and Moshe Looks, eds. 2011. *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings.* Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* New Haven, CT: Yale University Press.

Searle, John R. 1992. *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.

Shachtman, Noah. 2007. "Robot Cannon Kills 9, Wounds 14." *Wired,* October 18. http://www.wired.com/dangerroom/2007/10/robot-cannon-ki/.

Shulman, Carl. 2009. "Arms Control and Intelligence Explosions." Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

———. 2010a. *Omohundro's "Basic AI Drives" and Catastrophic Risks.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/BasicAIDrives.pdf.

———. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/WBE-Superorgs.pdf.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. 2009a. "Machine Ethics and Superintelligence." In Reynolds and Cassinelli 2009, 95–97.

———. 2009b. "Which Consequentialism? Machine Ethics and Moral Divergence." In Reynolds and Cassinelli 2009, 23–25.

Shulman, Carl, and Anders Sandberg. 2010. "Implications of a Software-Limited Singularity." In Mainzer 2010.

Smith, Michael. 2009. "Desires, Values, Reasons, and the Dualism of Practical Reason." *Ratio* 22 (1): 98–125. doi:10.1111/j.1467-9329.2008.00420.x.

Snaider, Javier, Ryan Mccall, and Stan Franklin. 2011. "The LIDA Framework as a General Tool for AGI." In Schmidhuber, Thórisson, and Looks 2011, 133–142.

Sobel, David. 1994. "Full Information Accounts of Well-Being." *Ethics* 104 (4): 784–810. http://www.jstor.org/stable/2382218.

Sobolewski, Matthias. 2012. "German Cabinet to Agree Tougher Rules on High-Frequency Trading." *Reuters,* September 25. Accessed December 31, 2012. http://in.reuters.com/article/2012/09/25/germany-bourse-rules-idINL5E8KP8BK20120925.

Solomonoff, Ray J. 1985. "The Time Scale of Artificial Intelligence: Reflections on Social Effects." *Human Systems Management* 5:149–153.

Solum, Lawrence B. 1992. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70:1231–1287. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1108671.

Sotala, Kaj. 2012. "Advantages of Artificial Intelligences, Uploads, and Digital Minds." *International Journal of Machine Consciousness* 4 (1): 275–291. doi:10.1142/S1793843012400161.

Sotala, Kaj, and Harri Valpola. 2012. "Coalescing Minds: Brain Uploading-Related Group Mind Scenarios." *International Journal of Machine Consciousness* 4 (1): 293–312. doi:10.1142/S1793843012400173.

Spears, Diana F. 2006. "Assuring the Behavior of Adaptive Agents." In *Agent Technology from a Formal Perspective,* edited by 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.

Stahl, Bernd Carsten. 2002. "Can a Computer Adhere to the Categorical Imperative? A Contemplation of the Limits of Transcendental Ethics in IT." In, edited by Iva Smit and George E. Lasker, 13–18. Vol. 1. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Staniford, Stuart, Vern Paxson, and Nicholas Weaver. 2002. "How to 0wn the Internet in Your Spare Time." In *Proceedings of the 11th USENIX Security Symposium,* edited by Dan Boneh, 149–167. Berkeley, CA: USENIX. http://www.icir.org/vern/papers/cdc-usenix-sec02/.

Steunebrink, Bas R., and Jürgen Schmidhuber. 2011. "A Family of Gödel Machine Implementations." In Schmidhuber, Thórisson, and Looks 2011, 275–280.

Suber, Peter. 2002. "Saving Machines from Themselves: The Ethics of Deep Self-Modification." Accessed December 31, 2012. http://www.earlham.edu/~peters/writing/selfmod.htm.

Sullins, John P. 2005. "Ethics and Artificial life: From Modeling to Moral Agents." *Ethics & Information Technology* 7 (3): 139–148. doi:10.1007/s10676-006-0003-5.

———. 2006. "When Is a Robot a Moral Agent?" *International Review of Information Ethics* 6:23–30.

Swan, Liz Stillwaggon, and Joshua Howard. 2012. "Digital immortality: Self or 0010110." *International Journal of Machine Consciousness* 4 (1): 245–256. doi:10.1142/S1793843012400148.

Sweeney, Latanya. 1997. "Weaving Technology and Policy Together to Maintain Confidentiality." *Journal of Law, Medicine & Ethics* 25 (2–3): 98–110. doi:10.1111/j.1748-720X.1997.tb01885.x.

Tanyi, Attila. 2006. "An Essay on the Desire-Based Reasons Model." PhD diss., Central European University. http://web.ceu.hu/polsci/dissertations/Attila_Tanyi.pdf.

Tarleton, Nick. 2010. *Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/CEV-MachineEthics.pdf.

Tenenbaum, Joshua B., Thomas L. Griffiths, and Charles Kemp. 2006. "Theory-Based Bayesian Models of Inductive Learning and Reasoning." In "Probabilistic Models of Cognition." Special issue, *Trends in Cognitive Sciences* 10 (7): 309–318. doi:10.1016/j.tics.2006.05.009.

Thomas, Michael S. C., and James L. McClelland. 2008. "Connectionist Models of Cognition." In *The Cambridge Handbook of Computational Psychology,* edited by Ron Sun, 23–58. Cambridge Handbooks in Psychology. New York: Cambridge University Press.

Trope, Yaacov, and Nira Liberman. 2010. "Construal-level Theory of Psychological Distance." *Psychological Review* 117 (2): 440–463. doi:10.1037/a0018963.

Turing, A. M. 1951. "Intelligent Machinery, A Heretical Theory." A lecture given to '51 Society' at Manchester.

Turney, Peter. 1991. "Controlling Super-Intelligent Machines." *Canadian Artificial Intelligence,* July 27, 3–4, 12, 35.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211 (4481): 453–458. doi:10.1126/science.7455683.

Van Gelder, Timothy. 1995. "What Might Cognition Be, If Not Computation?" *Journal of Philosophy* 92 (7): 345–381. http://www.jstor.org/stable/2941061.

Van Kleef, Gerben A., Astrid C. Homan, Catrin Finkenauer, Seval Gundemir, and Eftychia Stamkou. 2011. "Breaking the Rules to Rise to Power: How Norm Violators Gain Power in the Eyes of Others." *Social Psychological and Personality Science* 2 (5): 500–507. doi:10.1177/1948550611398416.

Van Kleef, Gerben A., Christopher Oveis, Ilmo van der Löwe, Aleksandr LuoKogan, Jennifer Goetz, and Dacher Keltner. 2008. "Power, Distress, and Compassion: Turning a Blind Eye to the Suffering of Others." *Psychological Science* 19 (12): 1315–1322. doi:10.1111/j.1467-9280.2008.02241.x.

Verdoux, Philippe. 2010. "Risk Mysterianism and Cognitive Boosters." *Journal of Futures Studies* 15 (1): 1–20. Accessed February 2, 2013. http://www.jfs.tku.edu.tw/15-1/A01.pdf.

———. 2011. "Emerging Technologies and the Future of Philosophy." *Metaphilosophy* 42 (5): 682–707. doi:10.1111/j.1467-9973.2011.01715.x.

Versenyi, Laszlo. 1974. "Can Robots be Moral?" *Ethics* 84 (3): 248–259. http://www.jstor.org/stable/2379958.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace,* 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf.

Walker, Mark. 2008. "Human Extinction and Farsighted Universal Surveillance." Working Paper, September. Accessed December 31, 2012. http://www.nmsu.edu/~philos/documents/sept-2008-smart-dust-final.doc.

———. 2011. "Personal Identity and Uploading." *Journal of Evolution and Technology* 22 (1): 37–51. http://jetpress.org/v22/walker.htm.

Wallach, Wendell. 2010. "Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making." In "Robot Ethics and Human Ethics," edited by Anthony Beavers. Special issue, *Ethics and Information Technology* 12 (3): 243–250. doi:10.1007/s10676-010-9232-8.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press. doi:10.1093/acprof:oso/9780195374049.001.0001.

———. 2012. "Framing Robot Arms Control." *Ethics and Information Technology.* doi:10.1007/s10676-012-9303-0.

Wallach, Wendell, Colin Allen, and Stan Franklin. 2011. "Consciousness and Ethics: Artificially Conscious Moral Agents." *International Journal of Machine Consciousness* 3 (1): 177–192. doi:10.1142/S1793843011000674.

Wallach, Wendell, Colin Allen, and Iva Smit. 2008. "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties." In "Ethics and Artificial Agents." Special issue, *AI & Society* 22 (4): 565–582. doi:10.1007/s00146-007-0099-0.

Wallach, Wendell, Stan Franklin, and Colin Allen. 2010. "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents." *Topics in Cognitive Science* 2 (3): 454–485. doi:10.1111/j.1756-8765.2010.01095.x.

Wang, Pei. 2012. "Motivation Management in AGI Systems." In Bach, Goertzel, and Iklé 2012, 352–361.

Wang, Pei, Ben Goertzel, and Stan Franklin, eds. 2008. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference.* Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Warwick, Kevin. 1998. *In the Mind of the Machine: Breakthrough in Artificial Intelligence.* London: Arrow.

———. 2003. "Cyborg Morals, Cyborg Values, Cyborg Ethics." *Ethics and Information Technology* 5 (3): 131–137. doi:10.1023/B:ETIN.0000006870.65865.cf.

———. 2010. "Implications and Consequences of Robots with Biological Brains." *Ethics and Information Technology* 12 (3): 223–234. doi:10.1007/s10676-010-9218-6.

Waser, Mark R. 2008. "Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence." In *Biologically Inspired Cognitive Architectures: Papers from the AAAI Fall Symposium,* 195–200. Technical Report, FS-08-04. AAAI Press, Menlo Park, CA. http://www.aaai.org/Papers/Symposia/Fall/2008/FS-08-04/FS08-04-049.pdf.

———. 2009. "A Safe Ethical System for Intelligent Machines." In *Biologically Inspired Cognitive Architectures: Papers from the AAAI Fall Symposium,* edited by Alexei V. Samsonovich, 194–199. Technical Report, FS-09-01. AAAI Press, Menlo Park, CA. http://aaai.org/ocs/index.php/FSS/FSS09/paper/view/934.

———. 2011. "Rational Universal Benevolence: Simpler, Safer, and Wiser than 'Friendly AI.'" In Schmidhuber, Thórisson, and Looks 2011, 153–162.

Weatherson, Brian. 2005. "Should We Respond to Evil with Indifference?" *Philosophy and Phenomenological Research* 70 (3): 613–635. doi:10.1111/j.1933-1592.2005.tb00417.x.

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,* edited by Barbara Hayes-Roth and Richard E. Korf, 1042–1047. Menlo Park, CA: AAAI Press. http://www.aaai.org/Papers/AAAI/1994/AAAI94-160.pdf.

Weng, Yueh-Hsuan, Chien-Hsun Chen, and Chuen-Tsai Sun. 2008. "Safety Intelligence and Legal Machine Language: Do We Need the Three Laws of Robotics?" In *Service Robot Applications,* edited by Yoshihiko Takahashi. InTech. doi:10.5772/6057.

Weng, Yueh-Hsuan, Chien-Hsun Chen, and Chuen-Tsai Sun. 2009. "Toward the Human–Robot Coexistence Society: On Safety Intelligence for Next Generation Robots." *International Journal of Social Robotics* 1 (4): 267–282. doi:10.1007/s12369-009-0019-1.

Whitby, Blay. 1996. *Reflections on Artificial Intelligence: The Legal, Moral, and Ethical Dimensions*. Exeter, UK: Intellect Books.

Whitby, Blay, and Kane Oliver. 2000. "How to Avoid a Robot Takeover: Political and Ethical Choices in the Design and Introduction of Intelligent Artifacts." Paper presented at Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights at AISB-00, University of Birmingham, England. http://www.sussex.ac.uk/Users/blayw/BlayAISB00.html.

Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation." *Science* 131 (3410): 1355–1358. http://www.jstor.org/stable/1705998.

Wilson, Grant S. Forthcoming. "Minimizing Global Catastrophic and Existential Risks from Emerging Technologies Through International Law." Working Paper.

Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Belknap.

Wood, David Murakami, and Kirstie Ball, eds. 2006. *A Report on the Surveillance Society: For the Information Commissioner, by the Surveillance Studies Network*. Wilmslow, UK: Office of the Information Commissioner, September. http://www.ico.org.uk/about_us/research/~/media/documents/library/Data_Protection/Practical_application/SURVEILLANCE_SOCIETY_SUMMARY_06.ashx.

Yampolskiy, Roman V. 2012. "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.

———. 2013. "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach." In *Philosophy and Theory of Artificial Intelligence*, edited by Vincent C. Müller, 389–396. Vol. 5. Studies in Applied Philosophy, Epistemology and Rational Ethics. New York: Springer. doi:10.1007/978-3-642-31674-6_29.

Yampolskiy, Roman V., and Joshua Fox. 2012. "Safety Engineering for Artificial General Intelligence." *Topoi*. doi:10.1007/s11245-012-9128-9.

Yudkowsky, Eliezer. 1996. "Staring into the Singularity." Unpublished manuscript. Last revised May 27, 2001. http://yudkowsky.net/obsolete/singularity.html.

———. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute, San Francisco, CA, June 15. http://intelligence.org/files/CFAI.pdf.

———. 2004. *Coherent Extrapolated Volition*. The Singularity Institute, San Francisco, CA, May. http://intelligence.org/files/CEV.pdf.

———. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom and Ćirković 2008a, 308–345.

———. 2008b. "Hard Takeoff." *Less Wrong* (blog), December 2. http://lesswrong.com/lw/wf/hard_takeoff/.

———. 2009. "Value is Fragile." *Less Wrong* (blog), January 29. http://lesswrong.com/lw/y3/value_is_fragile/.

———. 2011. *Complex Value Systems are Required to Realize Valuable Futures.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/ComplexValues.pdf.

———. 2012. "Reply to Holden on 'Tool AI.'" *Less Wrong* (blog), June 12. http://lesswrong.com/lw/cze/reply_to_holden_on_tool_ai/.

Zimmerman, David. 2003. "Why Richard Brandt Does Not Need Cognitive Psychotherapy, and Other Glad News about Idealized Preference Theories in Meta-Ethics." *Journal of Value Inquiry* 37 (3): 373–394. doi:10.1023/B:INQU.0000013348.62494.55.