Chapter 2
Risks of the Journey to the Singularity

Kaj Sotala and Roman Yampolskiy

2.1 Introduction[1]

Many have argued that in the next twenty to one hundred years we will create artificial general intelligences [AGIs] (Baum et al. 2011; Sandberg and Bostrom 2011; Müller and Bostrom 2014).[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). This article, however, focuses on the suggestion

[1]This chapter is based on three earlier publications (Sotala and Yampolskiy 2015; Sotala and Yampolskiy 2013; Yampolskiy 2013).

[2]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.

K. Sotala
Foundational Research Institute, Basel, Switzerland
e-mail: kaj.sotala@foundational-research.org

R. Yampolskiy (✉)
University of Louisville, Louisville, USA
e-mail: roman.yampolskiy@louisville.edu

V. Callaghan et al. (eds.), *The Technological Singularity*,
The Frontiers Collection, DOI 10.1007/978-3-662-54033-6_2

that AGIs may come to act in ways not intended by their creators, and in this way pose a *catastrophic* (Bostrom and Ćirković 2008) or even an *existential* (Bostrom 2002) risk to humanity.[4]

2.2 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. This argument will be further elaborated on in the following sections.

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 (Hanson 2008; Bostrom 2014). 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; Chalmers 2010; Bostrom 2014). 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; Bostrom 2014).

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

2.2.1 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,

[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 2008). 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; Vinge 1993; Joy 2000; Yudkowsky 2008a; Bostrom 2014).

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 et al. 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 (Dahm 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 (Congress 2000).

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 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 et al. 2011; Müller and Bostrom 2014). Eventually it will make economic sense to automate all or nearly all jobs (Hanson 2008; Hall 2008). As AGIs will possess many advantages over humans (Sotala 2012; Muehlhauser and Salamon 2012a, b; Bostrom 2014), a greater and greater proportion of the workforce will consist of intelligent machines.

2.2.2 AGIs Might Harm Humans

AGIs might bestow overwhelming military, economic, or political power on the groups that control them (Bostrom 2002, 2014). For example, automation could lead to an ever-increasing transfer of wealth and power to the owners of the AGIs (Brynjolfsson and McAfee 2011). AGIs could also be used to develop advanced weapons and plans for military operations or political takeovers (Bostrom 2002). 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 and SEC 2010).[5] 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 et al. 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 et al. 2006; Anderson and Anderson 2011), concerned with creating AI systems designed to make appropriate moral choices. Compared to narrow-AI systems, AGIs will be even more autonomous and capable, and will thus require even more robust solutions for governing their behavior.[6]

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."

[5]On the less serious front, see http://www.michaeleisen.org/blog/?p=358 for an amusing example of automated trading going awry.

[6]In practice, there have been two separate communities doing research on automated moral decision-making (Muehlhauser and Helm 2012a, b; Allen and Wallach 2012; Shulman et al. 2009). The "AI risk" community has concentrated specifically on advanced AGIs (e.g. Yudkowsky 2008a; Bostrom 2014), while the "machine ethics" community typically has concentrated on more immediate applications for current-day AI (e.g. Wallach et al. 2008; Anderson and Anderson 2011). In this chapter, 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.

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 2011; Muehlhauser and Helm 2012a, b; Bostrom 2014).

Muehlhauser and Helm (2012a, b) 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 machine programmed to maximize the satisfaction of human (or sentient) preferences might 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 2012a, b), that the pursuit of pleasure is not the only human value, and that humans are often unaware of their own values.

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

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 failed to incorporate 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).[7]

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. 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 only some AGIs are safe. 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.

2.2.3 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 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

[7]Miller (2012) similarly notes 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. "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).

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.[8] 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. A soft takeoff might however present its own problems, such as there being a larger number of AGIs distributed throughout the economy, making it harder to contain an eventual takeoff.

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.[9]

2.2.3.1 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.[10]

If such trends continue, then by the time the software for AGI is invented there may be a *computing overhang*—an abundance of cheap hardware available for

[8]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.

[9]Bostrom (2014, chap. 3) discusses three kinds of superintelligence. A speed superintelligence "can do all that a human intellect can do, but much faster". A collective superintelligence is "a system composed of large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system". A quality superintelligence "is at least as fast as a human mind and vastly qualitatively smarter". These can be seen as roughly corresponding to the different kinds of hard takeoff scenarios. A speed explosion implies a speed superintelligence, an intelligence explosion a quality superintelligence, and a hardware overhang may lead to any combination of speed, collective, and quality superintelligence.

[10]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 June 2016, 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^{14} FLOPS (Top500 2016). 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.

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.[11] 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.[12]

2.2.3.2 Speed Explosion

Another possibility is a *speed explosion* (Solomonoff 1985; Yudkowsky 1996; Chalmers 2010), 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 and stop, but by that time AGIs might come to accomplish most tasks at far faster rates than humans, thereby achieving dominance. (In principle, the same process could also be achieved via improved software.)

[11]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.

[12]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 et al. 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 et al. 2002).

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.[13]

2.2.3.3 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,[14] such that the intelligence of humankind is quickly left far behind and the machines achieve dominance (Good 1965; Chalmers 2010; Muehlhauser and Salamon 2012a, b; Loosemore and Goertzel 2012; Bostrom 2014).

Yudkowsky (2008a, b) 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 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.

---

[13]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.

[14]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.

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 (2010). 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.[15]

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.

References

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

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

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.

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

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.

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

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

Bostrom, Nick. 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. 2014. Superintelligence: Paths, dangers, strategies. Oxford University Press.

Bostrom, Nick, and Milan M. Ćirković. 2008. "Introduction." In Bostrom, Nick, and Milan M. Ćirković, eds. Global Catastrophic Risks. New York: Oxford University Press., 1–30.

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.

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.

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. 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.

Congress, US. 2000. National Defense Authorization, Fiscal Year 2001, Pub. L. No. 106–398, 114 Stat. 1654.

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.

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.

Hall, John Storrs 2008. "Engineering Utopia." In Wang, Goertzel, and Franklin 2008, 460–467.

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

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

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.usrobotics.us/reports/CCC%20Report.pdf.

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

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.

Loosemore, Richard, and Ben Goertzel. 2012. "Why an Intelligence Explosion is Probable." In Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. Singularity Hypotheses: A Scientific and Philosophical Assessment. The Frontiers Collection. Berlin: Springer.

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

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 Measurement (IMW '02), 273–284. New York: ACM Press. doi:10.1145/637201.637244.

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

Muehlhauser, Luke, and Louie Helm. 2012. "The Singularity and Machine Ethics." In Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. Singularity Hypotheses: A Scientific and Philosophical Assessment. The Frontiers Collection. Berlin: Springer.

Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. Singularity Hypotheses: A Scientific and Philosophical Assessment. The Frontiers Collection. Berlin: Springer.

Müller, V. C., and Bostrom, N. 2014. Future progress in artificial intelligence: A survey of expert opinion. Fundamental Issues of Artificial Intelligence.

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/.

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

Omohundro, Stephen M. 2012. "Rational Artificial Intelligence for the Greater Good." In Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. Singularity Hypotheses: A Scientific and Philosophical Assessment. The Frontiers Collection. Berlin: Springer.

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

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.

Sandberg, Anders. 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.

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/wpcontent/uploads/brain-emulation-roadmap-report1.pdf.

Sandberg, Anders, and Nick Bostrom. 2011. Machine Intelligence Survey. Technical Report, 2011-1. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/reports/2011-1.pdf.

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

Shulman, Carl, and Anders Sandberg. 2010. "Implications of a Software-Limited Singularity." In Mainzer, Klaus, ed. ECAP10: VIII European Conference on Computing and Philosophy. Munich: Dr. Hut.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. 2009. "Machine Ethics and Superintelligence." In Reynolds, Carson, and Alvaro Cassinelli, eds. AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings, 95–97.

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

Sotala, Kaj, and Roman V. Yampolskiy. 2013. Responses to catastrophic AGI risk: a survey. Technical report 2013-2. Berkeley, CA: Machine Intelligence Research Institute.

Sotala, Kaj, and Roman V. Yampolskiy. 2015. Responses to catastrophic AGI risk: a survey. Physica Scripta, 90(1), 018001.

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

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/.

Top500.org. 2016. Top500 list – June 2016. https://www.top500.org/list/2016/06/.

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.

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.

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.

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

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

Yampolskiy, Roman V. 2013. What to Do with the Singularity Paradox? Studies in Applied Philosophy, Epistemology and Rational Ethics vol 5, pp. 397–413. Springer Berlin Heidelberg.

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

Yudkowsky, Eliezer. 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.

Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom, Nick, and Milan M. Ćirković, eds. Global Catastrophic Risks. New York: Oxford University Press., 308–345.

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

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

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

http://www.springer.com/978-3-662-54031-2

The Technological Singularity
Managing the Journey
Callaghan, V.; Miller, J.; Yampolskiy, R.; Armstrong, S. (Eds.)
2017, XII, 261 p. 11 illus., 8 illus. in color., Hardcover
ISBN: 978-3-662-54031-2