FORESIGHT INSTITUTE

Cyber, Nano, and AGI Risks: Decentralized Approaches to Reducing Risks

by Christine Peterson, Mark S. Miller, and Allison Duettmann

The First International Colloquium on Catastrophic and Existential Risk,
UCLA B. John Garrick Institute for the Risk Sciences, 2017

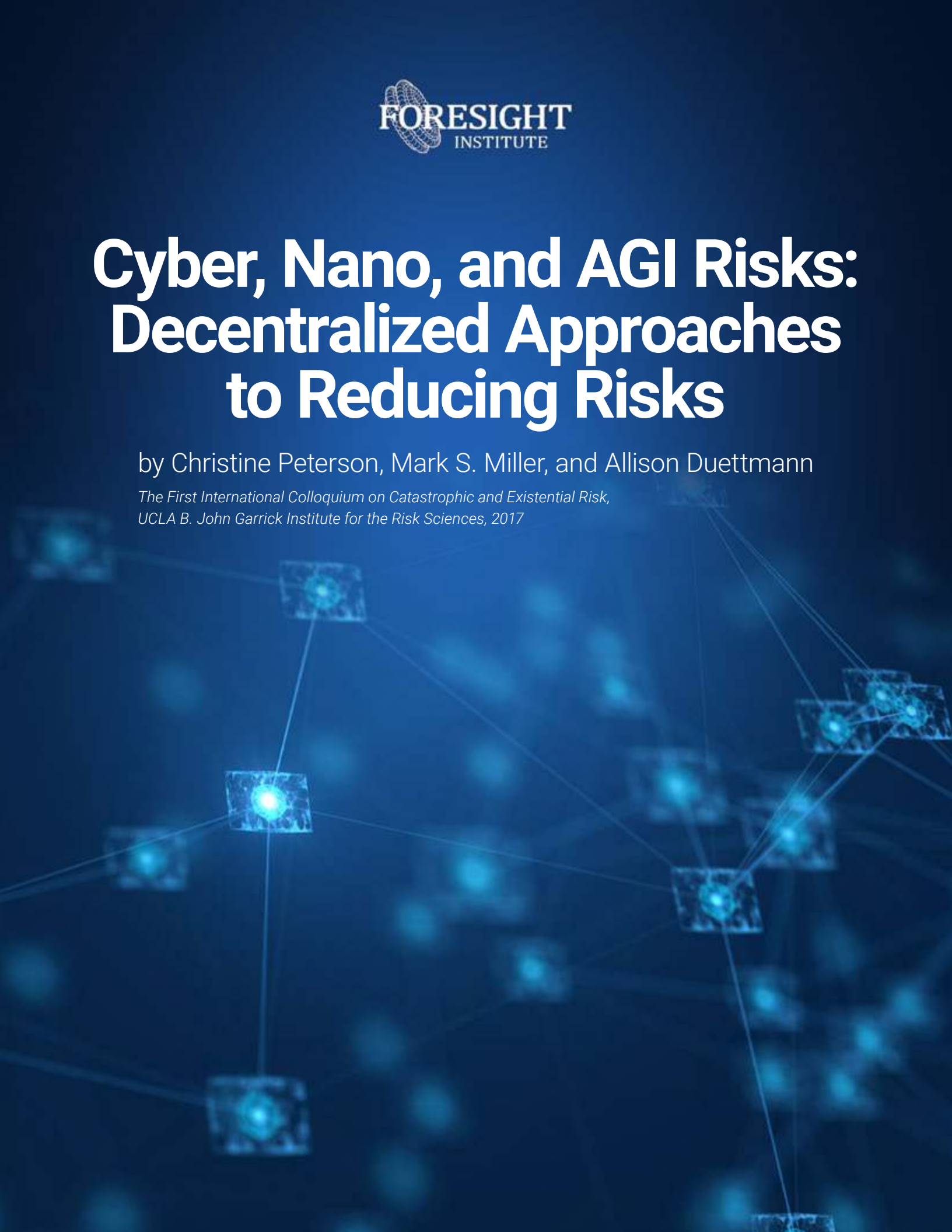

Table of Contents

Cyber, Nano, and AGI Risks: Decentralized Approaches to Reducing Risks

by Christine Peterson, Mark S. Miller, and Allison Duettmann
Foresight Institute, PO Box 61058, Palo Alto, California 94306 USA

"If men were angels, no government would be necessary. If angels were to govern men, neither external nor internal controls on government would be necessary. In framing a government which is to be administered by men over men, the great difficulty lies in this: you must first enable the government to control the governed; and in the next place oblige it to control itself."

—James Madison

Abstract

The aim of this paper, rather than attempting to present one coherent strategy for reducing existential risks, is to introduce a variety of possible options with the goal of broadening the discussion and inviting further investigation. Two themes appear throughout: (1) the proposed approaches for risk reduction attempt to avoid the dangers of centralized "solutions," and (2) cybersecurity is not treated as a separate risk. Instead, trustworthy cybersecurity is a prerequisite for the success of our proposed approaches to risk reduction.

Our focus is on proposing pathways for reducing risks from advanced nanotechnology and artificial general intelligence.

Nanotechnology can be divided into stages; even at the most advanced stage, society should be less worried about biology-style accidents than deliberate abuse. Development of nanotechnology weapons would not be detectable by current weapons monitoring techniques. An automated monitoring system, if based on sufficiently secure software foundations and physical arrangements, could serve as the basis for an arms control enforcement mechanism. Design needs to be open and decentralized to build the required public trust.

Civilization, taken as a whole, is already a superintelligence. It is vastly more intelligent than any individual, it is already composed of both human and machine intelligences, and its intelligence is already increasing at an exponentially accelerating rate. Civilization as a whole does not "want anything"; it has no utility function. But it does have a tropism—it tends to grow in certain directions. To the extent that its dominant dynamic emerges from non-coercive, non-violent, voluntary interactions, it is already shaped by human values and desires. It tends, imperfectly, to climb Pareto preferred paths. Society would address the value alignment problem more effectively by strengthening this dynamic rather than trying to replace it. No designed utility function would clearly serve human happiness better, and no replacement for civilization's dynamics is likely to be adopted anyway.

While still controversial, there is increasing concern that once artificial general intelligence fully surpasses the human level, human skills will have little or no economic value. The rate of economic growth will be extraordinary, but humans will have little comparative advantage. Will civilization still serve human preferences? When growth is extraordinary, so are returns to capital. The least-disruptive approach may be a one-time, gradual distribution of tradeable rights to as-yet unclaimed resources in space.

Introduction

The long term goal can be defined as human survival in the face of various existential risks, including those posed by both advanced nanotechnology and artificial intelligence that exceeds human intelligence. Many risk-oriented organizations, e.g., Center for the Study of Existential Risk (CSER), Future of Humanity Institute (FHI), and Future of Life Institute (FLI), and many future-directed philanthropists, e.g. Elon Musk, Jaan Tallinn (Tallinn, 2012), and Peter Thiel, rate artificial intelligence as one of the most important issues facing humanity, if not as the single most important issue to work on for enabling a positive future.

Definition Nanotechnology

Nanotechnology is a term used to describe precise control of the structure of matter. Four levels are defined, which describe the production of progressively more capable atomically-precise (AP) (and partially AP) nanosystems (Drexler, Pamlin, 2013). Level 1 includes chemical synthesis, nanomaterials synthesis, nanolithography, and biotechnology; Level 2 includes AP macromolecular self assembly and AP mechanical manipulation; Level 3 includes biomimetic or machine-based productive nanosystems; and Level 4 includes high-throughput atomically-precise manufacturing. While defensive military use scenarios have been described for the earlier stages of nanotechnology (Kosal, 2009), these levels have not been seen as involving catastrophic or existential risks. More substantial risks are expected at the highly advanced Level 4 stage, which is substantially longer-term (Drexler, 2007).

Definition AI

Artificial intelligence (AI) that exceeds human-level intelligence in all intellectual tasks is described using a variety of terms, including superintelligence (Bostrom, 2014), advanced AI (Russell, Norvig, 2009), smarter-than-human AI (Soares, 2016), strong AI (Muehlhauser, 2013), and Artificial General Intelligence (Goertzel, 2014). For simplicity, this paper will use the term Artificial General Intelligence (AGI hereafter). While AGI is sometimes used to describe an AI whose intelligence level is merely equal to that of a human, it is widely assumed that once AI reaches human level intelligence, it will soon surpass this level (Tallinn, 2012). With potential risks associated with the development of AGI becoming a greater concern, the new field of AI safety research, while still in its infancy, is growing fast and with high quality (e.g., OpenAI's launch in 2015, DeepMind's safety board established by Demis Hassabis in 2016). For an overview of different safety approaches represented in the field, see Mallah, 2017.

Biotechnology and cyber risks arrive earlier than, and are subsets of, nano and AGI risks

In the case of both advanced nanotechnology (physical technology) and AGI (software technology), there are closely-related concerns that arise earlier in time. Both advanced nanotechnology and biotechnology are based on systems of molecular machines, with biotechnology restricted to molecular machines similar to those which nature has discovered; thus it is theoretically a subset of the broader category of molecular machine systems referred to as nanotechnology. (Biotechnology is sometimes referred to as "nature's nanotechnology.") Similarly, the cyber attack risks of today are a small subset of what will be possible in the future from AGI. Eventually, cyber attacks will be performed by AGIs, while today's cyber attacks are often performed by today's existing superintelligences, such as corporations and nation states.

Even these earlier risks are very challenging: in fact they can seem harder to address than the long-term ones, because they seem more real and concrete. Biotechnology dangers and cyber attack dangers relate more closely to the current world, so it is easier to see why they are so challenging, while Level 4 nanotechnology and AGI are still relatively abstract, so it is harder to see why they are difficult. However, a world safe against a level 4 nano-attack would be a world already safe against biotechnology attack. Likewise, a world safe against AGI would also be a world already safe against cyber attack. Biotechnology dangers and cyber attack are risks worth addressing in regard to making the world a safer place, both for the practical value of solving these very real problems and the additional benefit of learning strategies and designing institutions applicable to the related longer-term challenges. For example, addressing cyber attack issues now will head off substantial concerns regarding cyber risk from the society's increasing vulnerability to attack on networked consumer products including threats to the control of self-driving cars (Kornwitz, 2017).

Focus on scenario of sophisticated attacker

With regard to biotechnology attack and cyber attack, there are two types of sophisticated attack scenarios: (1) by nation states that prepare attacks as weaponized systems for potential use in war, and (2) by the iteration of open attacks becoming more sophisticated over time, with the information needed for the attacks getting commoditized, eventually enabling a wide variety of players to engage in attacks. An example for the second attack scenario, which is also known as the "script kiddie problem" in the cyber world, is the Stuxnet virus. This virus was one of the most elite pieces of malware before its workings became understood, commoditized, and reused by less-advanced attackers who could not have constructed the attack originally. Therefore, in both cases, for current purposes we can consider sophisticated attackers as the primary threat.

1 Establishing the comparison

1.1. Biotechnology attack and Nanotechnology attack

Biotechnology Attack

Biotechnology attack and nuclear attack both have physical component to monitor

In contrast with cyber attacks, biotechnology attack at least has a physical component involved, although the physical aspects can be very small, e.g., a small lab with a small number of people. Because there is a physical component, there is in principle something physically observable in the process, which suggests that one possible place to look for a useful precedent is the response to nuclear proliferation. The current nuclear weapon situation is still very concerning, but humanity has survived since World War II without these weapons being used in battle again, and this is partly due to non-proliferation treaties backed by monitoring regimes, amongst other reasons.

Biotechnology attack prevention requires far higher level of monitoring than nuclear, but similar to nano

The challenge ahead is that the physical objects that must be monitored in order to detect hostile nuclear weapons activity are relatively large-scale and easy to verify—with useful data available even from satellites in space—compared to what will be needed to monitor for offensive biotechnology use. Monitoring styles can be divided into three broad categories:

- 1. Traditional top-down, Big Brother-style surveillance, which tends to lead to abuse;
- 2. A symmetric system of surveillance plus sousveillance (upward-looking monitoring), advocated in The Transparent Society , which would be destructive of privacy but could hold abuses in check (Brin, 1998); and
- 3. A decentralized automatic network of agents with "confinement," in which information is not revealed to humans unless clear, pre-agreed, tripwire criteria are triggered; this would be difficult to implement but avoids abuses of category 1 and the loss of privacyfrom category 2.

An early experiment with category 3 was made by William Binney, at that time a senior official at the NSA (U.S. National Security Agency), who initiated a monitoring system called Thinthread to perform a legal, Constitutional version of surveillance based on filtering and automatic encryption. Data on individuals was only unencrypted "if a judge found probable cause to believe the target was connected with serious crime, including terrorism." Unfortunately, the program was cancelled just before 9/11 and eventually replaced by a similar system without the filtering and encryption protections (O'Cleirigh, 2015).

The degree of monitoring required to prevent biotechnology attack would be comparable to a level of monitoring corresponding to a pervasive surveillance state. While finding an acceptable and working level of monitoring to detect hostile biotechnology would be very challenging (Omohundro, 2014), it is very much like the level of monitoring required to detect hostile nanotechnology weaponization, so it is at least very similar with regard to problem domain. Since both involve systems of molecular machines, both require verification at the molecular level: a daunting challenge.

As societal transparency and surveillance have been increasing over time, it has become increasingly difficult for independent third parties such as terrorists to hide a secret research program to develop biotechnology weapons. However, the ongoing global illegal drug trade demonstrates that the level of transparency and surveillance now in place is not effective at preventing even a large-scale illegal societal dynamic. The very substantial financial flows connected with the illegal drug trade constitute an extra-legal institutional framework operating in secret. Unauthorized bioweapons labs could take advantage of these same extra-legal mechanisms; the illegal drug trade in this sense preserves areas of secrecy usable by non-governmental weapons efforts. For decades, governmental programs to address this issue have involved attempting to reduce the illegal drug flow, with marginal success and high costs in terms of corruption, similar to that seen from alcohol prohibition in the U.S. during the last century. The question arises: if drugs were legalized, how much of this extra-legal financial institutional framework would survive? If society becomes sufficiently concerned about terrorists developing bioweapons, this still-controversial approach to undermining terrorism may be tried.

Level 4 Nanotechnology attack

Desirability of Quantitative Risk Assessment

Currently, Level 4 nanotechnology remains decades in the future. All of the other risks explored here are more urgent. But as with other technological risks, it is desirable to carry out a Quantitative Risk Assessment (QRA) as early as possible. This process, even when done very early in a risk field, helps clarify areas needing further work. For Level 4 nanotechnology, "it can sort out some of the more predictable risks" (Garrick, 2008).

Nanotechnology security entails biotechnology security; both require high-level monitoring

The reason why a world that is safe against nanotechnology attack is one that is already safe against biotechnology attack is that all of the attack vectors that are concerning for biotechnology are forms of attack that nanotechnology could engage in (i.e., highly advanced nanotechnology could engage in a number of other forms of attack, but it certainly includes all of those biotechnology attacks). So to defend against such nanotechnology attacks, one has to be able to defend against biotechnology attacks. Likewise, if there is going to be the degree of monitoring that prevents those hostile nanotechnology attacks from happening in the first place, then there will need to be monitoring of activities in small-scale labs with small numbers of people performing small-scale manipulation of

generally widely deployed synthesis mechanisms that are otherwise general-purpose.

1.2. Cyber attack and AGI attack

AGI security requires widespread cybersecurity

In testimony before the U.S. Senate Subcommittee on Space, Science, and Competitiveness Committee on Commerce, Science, and Transportation, OpenAI Co-Founder and CTO Greg Brockman stated, "The Internet was built with security as an afterthought, rather than a core principle. We're still paying the cost for that today, with companies such as Target being hacked due to using insecure communication protocols. With AI, we should consider safety, security, and ethics as early as possible, and bake these into the technologies we develop."

Researchers have already noted the importance of cybersecurity for AGI, but in a different context than will be discussed here (e.g., Yampolskiy, 2016; Bostrom, 2017). In order for the world to be safe against AGI, why must it already be safe against cyber attack? Even if the first AGI is confined within an impenetrable virtual box, we should expect knowledge of how to build AGIs to proliferate rapidly. Some will grow AGIs independently and release them. When encountering systems that are vulnerable to possible attacks, AGIs will often be able to discover these vulnerabilities and invent attacks. AGIs in the wild will only be limited by enforcement mechanisms that other systems use to limit interactions to agreed rules. The integrity of these enforcement mechanisms relies on the security of their underlying platforms.

Cyber attack

Operating systems are the most vulnerable level currently

Computer systems have multiple levels of vulnerability to cyber attack, including hardware, firmware, operating systems, and users. Of these, currently the easiest pathway for attack are the operating systems, therefore our discussion starts with this most urgent vulnerability.

Problem for cybersecurity is social constraint that could be overcome via genetic takeover

With regard to cyber attack, it is widely believed that improvements to safety are a matter of technological discovery or need for new research. However, most of the techniques required to build systems that are largely secured from cyber attack, with a few exceptions, have already been known since the 1960s and 1970s, e.g., capability-based security (Dennis, Van Horn 1966) . These techniques would actually be adequate if society could somehow reconstruct the computational world, from its beginning, on top of those techniques. The problem is that a multi-trillion dollar ecosystem is already built on the current insecurable foundations, and it is very difficult to get adoption for something that needs to rebuild the entire ecosystem from scratch. Thus, researchers have been exploring strategies to bridge from current systems to new secure ones, in what is in other contexts known as "genetic takeover," a term derived from biology (Cairns-Smith 1982). In a genetic takeover, the new system is grown within the existing system without directly competing with the existing system. The new system can coexist with the existing system, work in a world dominated by the existing system, and be competitive in that world. Once the new

system comes to be widespread enough, one can start to shift over to the new system, and the previous system eventually becomes obsolete.

A real-world analogy can be drawn with how society has adapted to earthquake risk. Faced with an installed base of existing, unsafe building infrastructure, instead of requiring an immediate demolition and reconstruction, building codes are written to require earthquake reinforcement to be done on a gradual basis as other renovations take place. Over time, the installed base becomes much safer.

Genetic takeover was possible in the past and there are some hopeful examples today

The computer industry has had genetic takeovers; for example, the move from mainframes to personal computers. The entire ecosystem of mainframe software rested on a few mainframe platforms, which thereby seemed to be permanently entrenched. The new personal computing ecosystem initial grew alongside, complementing rather than competing with the old one at first, but eventually displacing it. So the attempt to replace today's existing, entrenched software ecosystem is not hopeless; but it is very difficult. Currently, there are several promising efforts to grow securable infrastructure smoothly within the entrenched infrastructure, such as Capsicum (Watson 2012a), seL4 rehosting Linux (Nordholz 2011), Secure EcmaScript (Miller 2013), Sandstorm on Cap'n Proto (Filardo 2016), Monte (Simpson 2017), and CHERI (Watson 2012b). However, funding remains low in comparison to the urgency and importance of the challenge, and none of these projects has yet achieved widespread industry adoption.

The adoption barrier is often ignored, but critical to success and hard to overcome

This adoption barrier to making the world a safer place is ignored in most abstract discussions of AGI and nanotechnology attack, perhaps because one imagines that once humanity is faced with these dangers urgently, society will do what needs to be done. If there's a known technological solution for dealing with the dangers, it is natural to assume those most concerned will be able to get a majority to build, adopt, and deploy these technological solutions fast enough to avert disaster. In the case of massive cyber attacks, one would hope that government and industry would invest in rebuilding infrastructure on more securable bases. However, after seeing how weakly the world has reacted to cyber attacks that reveal massive vulnerabilities, this now appears to be unrealistic wishful thinking. The more likely reaction to the panic following a major breach will be to direct even more effort into entrenched techniques that do not and cannot work, because those are seen as recognized best practices. Techniques that actually could work will be seen as experimental and outside established best practices, best avoided in an emergency.

U.S. electric grid highly vulnerable to cyber attack today

As an example of a serious attack that could happen at the present time, the U.S. electric grid is vulnerable today to cyber attack (McLarty, Ridge, 2014), with damage estimates by Lloyd's ranging up to $1 trillion (Rashid, 2015). Damage to the electric grid via cyber attack can include physical as well as software damage, and would take months (arguably, years) to repair, leaving an entire multi-state region without power. Lloyd's, as an insurance company, focused on estimating financial damages rather than fatalities. While plans have been made at the federal level in the U.S., they were prepared under a previous Administration, and it is as yet unclear whether these or similar plans will be carried out (Executive Office of the President, 2016). Of the recommendations in this paper, making capability-based upgrades to grid software is by far the most urgent.

2 Implementing a Safety Approach

2.1. Safety against Biotechnology and Nanotechnology attacks

Monitoring and multilateral deployment

Advanced nanotechnology can and likely will be simulated well before it can be implemented; this has been termed "exploratory engineering" (Drexler, 1988). This time gap between knowing what is buildable and carrying out the actual construction creates a possible strategy to increase safety. This ability could combine with open source approaches; as examples today there is an active world of open source activities including impressive efforts such as the OpenWorm Project (Szigeti, 2014). A deployed defense system that could actually defend against a nanotechnology attack—which is much more achievable than attempting to prevent the attack to occur—would consist of a deployed fabric of systems that could detect and react based on trustworthy mechanisms; this proposal has been termed an "active shield" (Drexler, 1986).

Such a system would need to be based on both a high level of computer security and the decentralized form of monitoring described in section 1.1 above: a mutually-watching system of watchers. This complete system of decentralized defensibility, once deployed, would create and maintain an effective monopoly of force, enforcing rules of voluntarism and taking immediate physical action against malfunctioning watchers, to end noncompliance at early stages. Thus the goal of the system would be mutually assured survival, rather than mutually assured destruction, or even the mutual deterrence enforced by the threat of using nuclear weapons. In order for such a system to be considered trustworthy, it would need to be designed in an open source, open manner, and be on record as requiring a simultaneous multilateral release of deployment when such deployment eventually becomes possible. This ability to model systems well before actual construction is feasible creates a potentially useful time gap: a window in which it is possible to "design-ahead" (Drexler, 1986).

The design-ahead window

The design-ahead window, however, is an opportunity that seems unlikely to be successfully exploited in a safety effort. Even in a best-case scenario—a system is designed that, if deployed, would monitor for offensive use and take action to prevent that use—the danger remains that one side might get to deployability before its competitors and decide to carry out a first-strike attack.

Hobbesian Trap

Given the uncertainties involved in conflict, it would appear that all parties have a lot to gain from simultaneous multilateral deployment of a mutual defense system. Unfortunately, the technical designs resulting from sophisticated design-ahead also create a first-strike instability. This results in a "Hobbesian Trap" (Pinker, 2011), such that even if no party involved desires to start a conflict, the fear that another party would do so gives an incentive to perform a first strike. We see no simple answer to this challenge.

A multilateral deployment is the scenario that, if it can be arranged, would be the most trustworthy, given that it would require the least degree of trust in the non-corruption of any one institution.

2.2. Safety against AGI and cyber attacks

AGI

Dominant AI arrival scenarios

There is a particular safety scenario of AGI discussed specifically in the circles around Nick Bostrom (Bostrom, 2014) and those around Eliezer Yudkowsky (Yudkowsky, 2015a) that has become sufficiently dominant that it is worth explicitly contrasting with another perspective on the issue. The following is simplified and mostly focused on Nick Bostrom's scenario as outlined in his book Superintelligence: Paths, Dangers, And Strategies (Bostrom, 2014). Bostrom considers two scenarios for AGI ramping up: the slow takeoff scenario and the hard takeoff (fast) scenario.

Slow takeoff scenario

Slow takeoff scenario would be safer but is less likely

Let us first consider Bostrom's slow takeoff scenario. In some sense this is a straw man, because Bostrom believes that while a slow takeoff scenario would be safer, a hard takeoff scenario is more likely and more dangerous, thus more worthy of concern. We agree on this point, but discuss it first for reasons that will become clear. In a slow takeoff, AGI gradually emerges in a likely naturally multilateral environment. Consider a scenario in which the slow takeoff is happening in a world in which secure computing technologies—techniques such as capability-based security—have become a worldwide general adoption success, so that the world has become generally much safer against cyberattacks.

Civilization as relevant superintelligence

For this scenario, one of the most relevant observations is that civilization as a whole is already a superintelligence, composed of both human and machine intelligences, serving a great variety of different interests. Granted, as machines become more intelligent, the fraction of the intelligence of civilization contributed by machine intelligence will come to be greater than the fraction contributed by human intelligence. However, in some sense this is irrelevant, because the greater intelligence is the intelligence of civilization as a whole, so we can consider that to be the relevant superintelligence. While corporations, industries, and even nation states do not meet some of the criteria that are sometimes assumed for the idealized portrayal of superintelligence (e.g., they are limited by human speed on some non-parallelizable tasks, Yudkowsky, 2016a), the set of criteria they do fulfill is sufficient to merit describing them in that way with regard to possible risks.

Just as the intelligence of humans is often judged by their ability to achieve certain goals set by an intelligence test, one could measure society's intelligence by its ability to achieve the goals set by individuals using resources provided for this purpose. (Miller, Drexler, 1988) suggest this thought experiment: "One can imagine putting a person or an ecosystem in a box and then presenting problems and contingent rewards through a window in the box. A box full of algae and fish will 'solve' a certain narrow set of problems (such as converting light into chemical energy) and will typically pay little attention to the reward. A box containing an intelligent person will solve a different, broader range of problems. A box containing, say, an industrial civilization (with access to algae, fish and Bell Labs) will solve a vastly greater range of problems. This ability to solve externally posed problems can be taken as a measure of that ecosystem's 'intelligence' ".

Thus, in any slow takeoff scenario, in which AGI is gradually emerging, the intelligence of civilization is the superintelligence that is relevant.

Civilization as networks of entities making requests of each other

Civilization as a whole is largely composed of networks of entities making requests of other entities (Miller, Tulloh, 2016). Some of those entities are humans, some are software, and in this scenario some of those software entities are machine intelligences. The making of requests consists primarily of the mutually voluntary interaction of the party making the request and another party responding to the request. The response to the request might not be to serve the best interests of the request-making entity. However, human institutions, having evolved over many thousands of years, tend to shape interactions to be mutually voluntary and in the interests of both parties.

This definition resembles Minsky's societal definition of intelligence in which adaptive intelligence arises from a system being conflicted, rather than perfectly aligned. For humans, Minsky defends the multiple self view in which "a part of me wants this, a part of me wants that" (Minsky,1985); because humans that are guided only by hunger will soon die, it is only the interaction of hunger and other desires (pain avoidance, etc.) that enables the organism to survive. Civilization as a whole is the most complex known system of adaptive intelligence with conflicted parts, thus the relevant superintelligence.

Civilization encourages voluntary interactions

Civilization emerges from voluntary and involuntary interactions between individuals, with the balance continuing to shift towards the voluntary (Pinker 2011). Voluntary interactions happen when all participants expect to benefit,

or they would not participate. A Pareto preferred change makes at least someone better off and no one worse off is (Freudenberg et al .,1991). Voluntary interactions tend, imperfectly, to move the world in Pareto preferred directions—to benefit their participants without involuntarily harming non-participants. Thousands of years of evolution of norms, laws, and institutional frameworks enable humanity to arrange ever more complex patterns of cooperation. Civilization is thus, imperfectly, largely shaped by human preferences already. It is not that civilization has a utility function, but it has a tropism. Civilizations tends, imperfectly, to grow in Pareto preferred directions. Civilization is an entrenched working system that is already superintelligent and already serves human interests.

Civilization as relevant superintelligence that serves human interests

Imagining that a new, better system can be designed to take over the world and displace this entrenched system of civilization is rather unrealistic. Instead, the goal should be to amplify the existing process of civilization and to defend it, to increase the likelihood that it is not displaced. If the current system is entirely displaced, it seems unlikely that a new system with a more beneficial utility function would actually be implemented. Human effort would be better invested in working to prevent any such unitary revolutions, because it appears unlikely that their result will serve the interests of massive numbers of people.

Hard takeoff scenario

Hard takeoff scenario involves sudden unitary takeover

Bostrom's main concern regards the prospect of a hard (i.e., sudden) takeoff, in which one particular AGI instance reaches AGI first, performs a strategic takeover, and pursues its utility function. According to Bostrom, the most important strategy that humanity can use to make AI safe in that scenario, apart from setting up the initial conditions correctly, is to shape the AI's utility function so that it serves human interests, by selecting the right top-level goal. Bostrom states that "our entire future may hinge on how we solve these problems" (Bostrom, 2003).

Prevent hard takeoff using the technological knowledge that would make it possible

In the case of Bostrom's hard takeoff scenario, the AGI would displace human civilization as the overall framework of relevance for intelligence and come to dominate the world in a sudden manner. We argue that to the extent that this is the concern, but it is believed that humanity will have the ability to constrain what the AGI does (e.g., by giving it the correct top-level goal), then any abilities that humans have to constrain such an AGI should instead focus on setting up an alternative, decentralized distribution of AGIs with a system of checks and balances, rather than trying to constrain one AGI to act in human interests.

Make superintelligence part of fabric of civilization

If humans are in a position to design what the initial breakout technology is able to do, then they should also be in a position to prevent it from performing a unitary strategic takeover. Instead, our efforts can focus on directing the technological ability that the breakthrough represents to itself become widely deployed as non-coercive entities in the world. These non-coercive entities can then take part as interactive agents in the fabric of civilization, deployed by different parties simultaneously to serve many different ends. This proposal has some similarity to Drexler's

technical proposal to distill superintelligent machine intelligence to apply only to specific problem domains, while avoiding the creation of one agent that has general intelligence, at least until a solution to AI safety is reached (Drexler, 2015). This would enable the use of many targeted, general-but-restricted AIs without requiring or entailing a unified AGI (Drexler, 2017).

The safety of civilization rests on its lack of a utility function, i.e., it is a negotiated compromise using an institutional framework that accommodates a great diversity of different ends. Thus, the safety relies on the fact that the simultaneous deployment of many instantiations of such a superintelligence would occur with the many instantiations serving many different ends, and no one entity being in a position to dominate. Additionally, most of those goals should be best served by cooperating with other entities, in extensions of the cooperative framework of civilization, just as most human goals are today. This game-theory-style approach has been described more generally: "The examples of memes controlling memes and of institutions controlling institutions also suggest that AI systems can control AI systems" (Drexler, 1986).

Civilization is already tested against AGIs

As mentioned earlier, civilization has already demonstrated its accommodation of superintelligences, in that large institutions themselves are already superintelligences with diverse interests that are interacting in a mostly mutually voluntary fashion. Thus, the stability of civilization has not only been tested by humans, it has also been tested by multiple interacting superintelligences, and has survived largely successfully.

Bostrom places hard moral philosophy between humans and safety

A difficulty with the approach pursued by Bostrom, Yudkowsky, and others (Armstrong, 2014) is that in attempting to construct a powerful entity that acts in human interests, it is necessary to ask some deep philosophical questions about what is it that humans want or should want and assumes that this question can be answered satisfyingly by the designers (Duettmann, 2014). For instance Bostrom notes the danger of the Paperclip Maximizer Scenario, in which humans want to give the AGI an apparently peaceful goal such as maximizing paperclips, and the AGI executes the literal command and maximizes paperclips by converting most of the matter in the solar system (humans included) into paperclips (Bostrom, 2003). While the types of concerns expressed in this thought experiment are valid, these are deep philosophical questions about what humans really want or even, as Yudkowsky states, what humans would really want "if we knew more, thought faster, were more the people we wished we were, and had grown up farther together" (Yudkowsky, 2004).

Yudkowsky views the most important issue regarding AGI as " constructing superintelligences that want outcomes that are high-value, normative, beneficial for intelligent life over the long run; outcomes that are, for lack of a better short phrase, 'good.'" (Yudkowsky, 2015b). Even Yudkowsky's less ambitious suggestion to construct a "Task AI", that is less sovereign than a full AGI, still relies on constructing partial normative theories. Yudkowsky calls this suggestion "insanely difficult" (Yudkowsky, 2016b). We agree . Bostrom refers to these as value-loading problems and acknowledges that AI safety must be "philosophy with a deadline" because focusing on human philosophical exploration into areas such as metaphysics doesn't contribute to solving the value-loading problems (Bostrom, 2014). However, even contemplating the extremely complicated value-loading problems, and attempting to construct the perfect goal, might well result in a completely different outcome, because the technological breakthrough will occur before philosophers have arrived at any satisfying answers to these questions. It is likely that human designers simply "do not possess the full wisdom needed to implement and grow a flawlessly benevolent intelligence" (Steunebrink

et al., 2015), not least because the AI research community lacks the diversity required to represent a wide enough range of different interests well (Li, 2016).

Rather than positioning the answers to philosophical questions that have caused disagreement for thousands of years between humanity and safety, it seems advisable to construct potential solutions which avoid moral questions that are this unanswerable.

Avoiding Benevolent Dictator scenario

A unitary takeover, whether fast or slow, is a "Benevolent Dictator" scenario at best. For much of human history, the central question of political philosophy was "Who should rule?". Political philosophy finally advanced once society realized that this was the wrong question, and to question instead whether there must be a unitary ruler (Popper 1945). Although Yudkowsky and Bostrom seek to construct the perfect dictator rather than to find one, this quest does recapitulate many of the problems of this old framing.

Ideal safety strategies would work despite uncertainty in timeframe

In current discussions of AGI safety, attempts are often made to estimate a median, average, or otherwise most-expected timeframe for the arrival of the technology. However, timeframe estimates vary by at least one order of magnitude and sometimes more, from relatively near-term (Kurzweil, 2012) to very long-term (Ng, 2015). Tools such as prediction markets (Hanson, 2003) and reputation-based prediction sites such as Metaculus (Aguirre, 2017) may be of some help in clarifying timeframes, but currently uncertainty remains high. In this situation, attempting to make a useful estimate of expected timing is overly optimistic; the error bars are too large. It appears advisable to develop AGI safety strategies that are robust against both early-arrival scenarios and late-arrival scenarios. A similar point has been made about the timeframe of risks from advanced nanotechnology (Drexler, 1986).

Implementing Secure Computing

The world is not yet hostile enough to incentivize secure computing systems today

The AGI safety and cyber attack safety strategies above require secure computing infrastructure. The adoption of secure computing is being delayed because the overall software ecosystem is not currently "hostile enough," i.e., companies and institutions can be too successful when they build systems that are very high-quality on many dimensions but are implemented in architectures that are insecurable.

Small projects can now free-ride on larger projects' being more attractive targets

In today's world in which primarily large-scale, entrenched software projects get attacked, most damage to early-stage software projects is due to dangers other than security. Therefore, for most early projects, investing in costly security is less important than investing in other areas, e.g., assembling the product and receiving feedback from user experience. Additionally, when hiring employees, a small company considers the additional value of the person to the project, so with regard to security, companies generally minimize the education burden that their team has to take on by following what are widely viewed as current best practices, rather than more unusual (and more secure) techniques. Consider the maxim that "to escape from a bear, one doesn't have to outrun the bear, but merely the

other guys"; if a small project engages in the same allegedly best practices as bigger projects, it can escape attack because other projects are bigger targets. By the time the small project becomes large, it would then be a serious target, but by that point it has enough capital to manage the security problem without truly fixing it. Currently, all large corporations are managing their pervasive insecurities rather than fixing them.

Current system is only sustainable because attacks are not very sophisticated yet

This situation is only survivable because nation-states are developing the most sophisticated attacks but not yet deploying them seriously. Additionally, the attacks that nation-states are developing are probably much less sophisticated than the attacks that the most advanced organizations could be engaging in by making better use of bleeding-edge early technologies combined with static analysis technologies. For instance, the strategies that are known from the Snowden revelations include gathering Zero-Day Attacks, i.e., entities wanting to take over others' computers accumulate Zero-Day Attacks, to prepare for a future day when that entity will use them against those target computers owned by others (Wikileaks, 2013). However, rather than gathering known Zero-Day Attacks, one can imagine software that is able to analyze the software being attacked and find entirely new, previously unknown Zero-Day Attacks. Having the best state-of-the-art software for discovering vulnerabilities built into the deployed attacking system would enable the system to discover vulnerabilities and exploit them while it is in active contact with the target, rather than just launching built-in attacks against previously known vulnerabilities. This level of attack software is one that the currently entrenched architectures are not going to survive, and it is likely to precede AGI.

The launch of a sophisticated attack would make the world hostile enough to end fragile systems, but would also severely disrupt it

On the positive side, at the point that this higher level of attack gets deployed, the world's software ecosystem will become hostile enough that the relative safety through obscurity of smaller, earlier projects will end because now insecurable systems of all sizes will be punished early on. The downside of this situation is that it comes with the danger of widespread destruction of the existing software infrastructure. If a certain threshold of the world's installed software base is destroyed, it could be difficult to transition to a safer situation without having gone through a serious downturn in overall functionality of the world's computation systems, not to mention the world economy.

seL4 microkernel as example of code safe against attacks

Combining various state-of-the-art research has led to some impressive results at finding vulnerabilities in software targets. One example is research on combining Machine-Learning sophisticated AI with sophisticated static analysis of programs to find vulnerabilities (Brooks 2017). This level of sophistication is not accidentally going to be part of an attack system. However, if it is built in as part of an experiment run on a platform that is believed to be secure but that is not air-gapped (i.e., is not isolated from the internet), such an experiment would be very good at detecting flaws. The seL4 microkernel is our best example of an operating system kernel that seems to be secure, due to its formal proof of end-to-end security and its track record of having withstood a Red Team Attack (a full-scope, multilayered attack simulation) which no other software has withstood (Fisher, 2014). One hopeful development is increased funding of seL4 by the U.S. Department of Defense. Nevertheless, its security rests on some counterfactual assumptions, such as that the formal model of the underlying hardware is accurate.

A model of decision alignment

A model of software object security can be combined with a model of human-to-human security

Many complex systems can be described as networks of entities making requests of other entities. In economics, there are principal-agent relationships, in which a principal sends a request to an agent . The principal uses various techniques to try to align the decision of the agents with the interests of the principal to increase the likelihood that the request is fulfilled.

Economics, for instance, studies principal-agent relationships among humans and examines both hazards, such as divergent interests and asymmetric information, and techniques for addressing those hazards. Software engineers deal with principal-agent relationships among computational objects and examine hazards and techniques such as object design patterns. Human Computer Interaction (HCI) deals with human-object interactions and examines hazards and techniques, such as user confusion or request expressiveness.

The techniques principals use to align agent decisions with the principal's intent can be divided into six categories: Select agent (admission control), Inspect internals (static analysis), Allow actions (least authority), Explain request (abstraction design), Reward cooperation (incentives), Monitor effects (reputation feedback); see Table 1 (Miller, Tulloh 2016). A unified view that looks at the different techniques in relation to each other can provide important insights. Reasoning across both rows and columns of Table 1, and combining techniques, allows reaping the payoff of having different techniques reinforce each other. Thus, Table 1 is not simply about reasoning by analogy, but instead reasoning about a single integrated network spanning multiple systems.

	Human to Human	Human to Object	Object to Object
Select	Trademark Chain of custody	App stores White and black lists	Trusted developer Same origin
Inspect	Accounting controls	Trusted path URL bar	Types, Verification Open source eyeballs
Allow	Law, Contracts	App permissions Powerbox	Security Protection patterns
Explain	Language	User interface	Abstraction
Reward	Economics Incentive Alignment	Objective functions	Machine learning Agorics
Monitor	Reviews, Complaints Word of mouth	Bug reports	Contracts, Testing Backprop

Table 1: *The Elements of Decision Alignment*

For example, computer security (Allow actions) taken alone misses some differences among agent actions that cause harm to the principal, such as when the agent benefits from misbehavior (Reward cooperation). Instead, principal-agent arrangements can be designed such that each technique fills in for weaknesses in the others, creating greatly increased structural strength built out of individually breakable parts.

While perfect security might ultimately be unattainable (Yampolskiy, 2016), this approach has the possibility of delivering adequate security and is a great deal more secure than any of the insecurable security systems that are now widely in use. Moreover, it is not only applicable to today's computer security but also is independent of the intelligence of the agent and therefore can be applied to AGI safety as well.

The problem of supply chain risk

Formal security proofs rest on assumption that hardware is safe but it might not be

After insecure operating systems, supply chain risk is the hardest problem in attempting to ensure secure computation. The proof that a given hardware chip design is secure only helps if hardware which the software is run on is actually the hardware that was designed. This assumption sounds trivial but it may be false, because it is possible that the hardware includes a manufactured-in trap door. Based on the revelations about the U.S. National Security Agency (NSA) serving national security letters to software companies forcing them to disclose user information, it is possible, indeed likely, that the NSA has already served national security letters to hardware companies including Intel and AMD requiring them to install trap doors into their hardware, which the NSA can later choose to trigger (Gustin, 2014). Fearing billions of USD in profit losses after the revelations in 2014, IBM's President Weber was quick to point out in an open letter to clients that the hardware giant would not comply with such letters (Weber, 2014). However, the severe penalties associated with disobedience or disclosure should cause us to be skeptical. None of today's proofs of software security can defend against such trap doors.

Open source processor design as possibility to overcome trustworthiness issues of hardware

In the near term one can imagine a technology example that can be secure against those risks: a good open source processor design for which there is a proof of security comparable to the proof of security of the seL4 software. There are many open source processor designs that are sufficiently high performance that, when run on a field-programmable gate array (FPGA), can run fast enough to be practical for many applications. By combining these well-designed processors with a layout algorithm that randomizes layout decisions, the processor could be randomly laid out for each individual hardware instance. Given this randomized layout, there is no feasible corruption of the FPGA hardware that can escape notice under electron microscopes and that would also be able to successfully corrupt most instances of the processor.

Even if trustworthy processor is theoretically possible, it is likely too expensive

However, even if it was possible to build a secure processor, it would be hopelessly unadoptable. The current norm in secure software holds that if a software security mechanism costs a factor of 3% more than insecurable mechanisms, widespread adoption becomes very unlikely. The trustworthy hardware in the form of a secure FPGA described above would result in a factor of at least an order of magnitude in performance cost over producing chips the standard way, which renders its adoption unrealistic.

The blockchain ecosystem an approach for safety

Ethereum and blockchain evolving in hostile ecosystem

A counterexample to the difficulties expressed above is Ethereum's current approach. Both Bitcoin and Ethereum are evolving in an ecosystem that is already under the very hostile attack pressures described earlier in this paper. When insecurity leads to losses, the players have no other recourse to compensate. Systems that are not bulletproof will be killed early and visibly, and therefore these ecosystems remain populated only by bulletproof systems. The bulletproof security of these systems are an essential part of their value proposition.

Ethereum as virtual machine that is trustworthy

Regarding the problem of trustable hardware mentioned earlier, if Ethereum is a virtual machine, it is a factor of at least ten thousand times more costly in performance than the FPGA approach mentioned earlier, that was already too expensive to be adopted. Ethereum is trustworthy in the same sense that Bitcoin is trustworthy; Bitcoin is a payment system and Etherium is a general purpose virtual machine (CPU, memory, limited IO). Both are synthesized by cryptographic protocols and massive redundancy among their players, based on a blockchain—an agreed order of messages. In order for either to take action in an untrustworthy fashion, a supermajority of participants would have to perform actions that were visibly illegitimate.

The Ethereum and Bitcoin systems per se are holding up very well. The publicized attacks on these systems do not reveal weaknesses in these foundation, but rather in the participants—at two different abstraction levels. Bitcoin exchanges were hacked, with losses of several hundred million dollars, due to insecurity of the platforms used by the exchanges, not due to any flaw in the Bitcoin protocols. Ethereum, as a virtual machine, runs programs written by its users, such as the DAO (Decentralized Autonomous Organization) smart contract.

Example of DAO as bad software deployed on top of trustworthy Ethereum machine

The DAO was the first significant piece of software deployed by a commercial participant on Ethereum, and it was not bulletproof. While the software withstood initial code review, it should have been subject to (at least) more code review, or preferably a formal proof of correctness. As explained above, machine checked formal proofs of correctness can be and have been successfully performed on much larger and more complicated pieces of software such as seL4.

In the case of The DAO, once this insecure piece of software was deployed, hackers exploited a known bug, started diverting money, and successfully removed US$60 million worth of Ether. This provoked the Ethereum ecosystem to engage in a "hard fork," a deliberate change of software that created a new version of the Ethereum system in which the Ether was not stolen (Buterin, 2016). Resetting the system in this way was a serious compromise of the founding principles of these cryptographic smart contract systems, which is that they are permissionless, i.e., that "code is law." It established a terrible precedent that future actions within the systems may be overridden by retroactive fiat.

Blockchain ecosystem as hope for building something that is secure against cyber attacks

Despite this early misstep, we are optimistic that the universe of cryptographic smart contracts can be the beginning of an ecosystem in which projects can grow up under extraordinarily hostile conditions. Such projects are evolving with a degree of adversarial testing that can create the seeds for a system that can survive a magnitude of cyberattack that would destroy conventional software. If this type of secure system grows enough before the world is subject to such cyberattacks, then a successful genetic takeover scenario might be achieved.

Safety of the proposed system still relies on counterfactual assumption

Bostrom states that even if one has a truly secure system, an AGI is likely to be able to break out of it, because "even a 'fettered superintelligence,' that was running on secure hardware on an isolated computer that can communicate only via text interface, might be able to break out of its confinement by persuading its handlers to release it" (Bostrom, 2003). While we cannot rule out this possibility, this degree of human gullibility does not seem plausible to us. Perhaps some pre-AGI experiments could help quantify this issue.

The proposed system's formal safety is independent of attacker intelligence, so would remain safe not only under cyber attack but also under AGI

While these other threat vectors are problematic, it is important to emphasize that, to the degree to which these systems are formally secure, that security is independent of the intelligence of the attacker. Thus, if humanity succeeds at building systems before AGI that are actually secure, which can in principle be done, then those systems should remain formally secure under AGI. The formal security of systems such as seL4, and the adversarial testing carried out on smart contracts, is likely to create an ecosystem of software systems which are secure against AGIs, because the threshold that needs to be crossed to guarantee security can be crossed well before AGI is reached. (In fact, this level could have been crossed before reaching the level of current machine intelligence.) There is no prerequisite of one on the other.

Successful Constitutions as role models for a safe system

As a few examples of organizational arrangements that have had some long-term success at managing competing superintelligences, we can point to the Swiss Federal Constitution, the U.S. Constitution, and the (partly unwritten, but real) U.K. Constitution. Here we take the U.S. Constitution as an illustration, primarily due the authors' relative familiarity with it; later work should address a wider variety of successful arrangements.

Founding fathers were trying to create a Constitution that depessimizes

The originators of the U.S. Constitution, termed the Founding Fathers, were faced with a Bostrom-like nightmare of having to design a single institution that was going to be superintelligent, and that was composed of systems of people that individually want to take many actions that society would collectively not want any of them to do. However, these originators felt that they had no choice but to design this institution and attempt to create an architecture that was inherently constructed to maintain its integrity, not at being ideal but at avoiding being very seriously flawed. This strategy is generally known as depessimizing, rather than optimizing as advocated by most in the AI safety field. The worst-case scenarios of our future are extremely negative and numerous, so by simply avoiding the worst cases humanity would be doing extraordinarily well. Attempting to do better among the non-worst-case scenarios can be

viewed as a very minor issue compared to safety against the worst cases. Goertzel advocates this depessimizing approach for AGI work: "the most sensible medium-term goal for human society is to guide the advance of technology in a rational way that has reasonable odds of getting past the current phase of development without causing global annihilation or other horrible catastrophes" (Goertzel, 2015).

Success of Constitution as lending support to feasibility of building safe AGI

In the case of AGI, instead of attempting to build an optimal system, humanity should focus on not building a system that turns into a worst-case scenario. In the case of the U.S. Constitution, instead of attempting to design the Constitutions as a optimized utility-function that would serve everyone's interests, the originators' main objective was to avoid having it becoming a tyranny. It is extraordinary that this Constitution maintained most of its integrity of mechanism, as well as integrity of purpose, for its first 150 years and maintains much of this even today. It shows that this type of effort can succeed and is worth taking on.

AI safety is harder than Constitution because less familiar knowledge to build predictions on

AI safety is harder in the sense that when formulating the Constitution, the Founding Fathers could rely on their knowledge of human nature and the history of politics and human institutions. With regard to AI safety, there is less solid ground; however, the basic mechanism can be based on the above, in a more focused and less decentralized way. Just as future AGIs will dwarf current superintelligences with regard to intelligence, so are current superintelligences dwarfing the expectations of what the Founding Fathers imagined when designing the Constitution over two centuries ago. Nevertheless, the Constitution was only designed as a starting point, on which later, more intelligent agents could build, and it is still surprisingly relevant one industrial revolution later. Rather than inventing a safety approach from first principles, a useful approach could make use of the immense body of historic and cultural knowledge that can be relied on to ensure a more organic AGI world, similar to the one envisioned by Kurzweil: "Ultimately, the most important approach we can take to keep AI safe is to work on our human governance and social institutions. We are already a human- machine civilization. The best way to avoid destructive conflict in the future is to continue the advance of our social ideals, which has already greatly reduced violence" (Kurzweil, 2014)

Elon Musk also appears to favor a decentralized approach: "The important thing is that if we do get some sort of runaway algorithm, then the human AI collective can stop the runaway algorithm. But if there's a large, centralized AI that decides, then there's no stopping it" (Dowd, 2017). Musk, Thiel, Reid Hoffman, Sam Altman, and others have founded and pledged a total of $1 billion to OpenAI, a foundation with the purpose of developing and distributing AI widely as a safety strategy (Risley, 2015; OpenAI, 2017).

Co-Chair Sam Altman explains, " Just like humans protect against Dr. Evil by the fact that most humans are good, and the collective force of humanity can contain the bad elements, we think it's far more likely that many, many AIs, will work to stop the occasional bad actors than the idea that there is a single AI a billion times more powerful than anything else" (Levy, 2015). The Open AI approach has attracted a $30 million grant from the Open Philanthropy Project (Open Philanthropy Project, 2017). For guidance on the AI transition, Altman has looked to James Madison's notes on the Constitutional Convention (Friend, 2016).

Multilateralism and gridlock as important part of the system

Previously we mentioned civilization being very widely multilateral. In that sense, the evolved institutions of civilization are the result of this decentralized, ongoing negotiation among institutional frameworks having a very wide diversity of interests. Additionally, the Madison form of government was a perpetually explicitly renegotiated framework among these small number of divergent interests that were purposely put its opposition with each other, including division of power, checks and balances, and significant decentralization. Building the system to be in conflict with itself is a much more realistic strategy than to pursue building a unitary system that wants the right goals. While the checks and balances designed into such a system lead to a decrease in speed and efficiency, this is a positive tradeoff in exchange for a reduction in much more serious risks.

In addition to the UK, US, and Swiss constitutions, those attempting to design governance systems for AGI safety may find inspiration from (1) John Locke on institutional checks and balances, (2) John Adams on federal-state balance, based on his study of the United Provinces of the Netherlands, Switzerland, the Holy Roman Empire, and the Peloponnesian League confederation in ancient Greece, and (3) the later cases of the Canadian, Australian, postwar German, and postwar Japanese constitutions (Bennett, 2017).

3 Securing human interest in an AGI world

How to achieve human interests once AGI is reached

Regardless of the exact timeframe, if our civilization and technology continue to progress, AGI will ultimately be reached. As Sam Harris points out, the only reason why AGI would not be reached eventually will be that an even worse event occurs, which destroys technology or civilization before it reaches that state (Harris, 2016). To ensure that civilization still serves human interests when AGI is reached, we argue that humans need a claim to capital to be able to participate in exchange, and that a promising way for humans to obtain this claim to capital is through a one-time distribution of unclaimed resources, referred to as "Inheritance Day" (Drexler, 1986).

3.1. Providing humans with capital bargaining power

This section argues that to ensure civilization still serves human interests once AGI is reached, humans need capital to participate in exchange for two main reasons:

Civilization serves human interests as long as humans contribute either skill or capital in exchange

Earlier in the paper we argued that civilization tends to serve human interests, albeit imperfectly. However, the argument that the system serves human preferences depends on humans having something to offer in exchange, either proceeds from their capital or their skills. Historically, much of what humans had to offer in exchange was based on human skills. These skills were of two kinds: human mechanical skills (the ability of humans to perform actions physically) and mental skills. Currently, while it is still possible to earn income using their skills, many humans don't have capital.

Once AGI is reached human skills become irrelevant but capital has high returns

Machines have already displaced humans from being able to earn much income via direct contribution of human mechanical skill and will continue to do so (McKinsey, 2015). Human dexterity has a lot to contribute currently, but that is largely due to its coupling with human mental ability. Human mental ability in the abstract contributes a great deal today, but humanity should anticipate the day when a machine intelligence achieves general intelligence. As pointed out increasingly by tech entrepreneurs like Bill Gates or Elon Musk, once AGI is widely deployed, human skills will be outcompeted and have little or no economic value.

An economy that is so productive that human skill is irrelevant is also an economy that grows extraordinarily quickly, similar to Robin Hanson's description of a economy in which the GDP is doubling in weeks (Hanson, 2014). Any economy growing at this rate offers extraordinary returns to capital. Capital is defined here as the ownership of resources, where resources are both physical objects as well as ownership of created abstract rights (i.e., corporate stock) that are part of the fabric of the civilization. This capital itself can become investments which the economy would reward extraordinarily.

If humanity enters into the transition to AGI with insufficient preparation, much of humanity will have no capital and their skills would be irrelevant. One strategy to ensure that the dynamic of civilization still contributes to human well-being once human skills are irrelevant is to arrange that most human beings have some capital claim that they can continue to trade on, get capital returns on, and live well.

3.2. Inheritance Day as strategy to provide capital claim

Inheritance Day as strategy to provide capital claim

Since humans need a claim to capital to ensure that civilization still serves their interests, this section deals with a potentially useful way of granting capital claims to humans. We argue that one possible strategy to provide humans with a capital claim is via a strategy called "Inheritance Day."

One-time distribution can provide humans with a capital claim

Capital claims can be assigned to individuals either by redistributing existing capital claims or distributing new capital claims. Currently, redistribution is the most common strategy to assigning capital claims to individuals in need. However, redistribution leads to political opposition, because it involves giving new beneficiaries a claim to capital by taking it away from the previous owners. Continual redistribution also appears to reward high reproductive rates, leading to additional opposition. To avoid this political opposition, rather than redistributing capital that has already been claimed, society could distribute capital that has not been claimed yet. While the great majority of the Earth's land and much of its undersea area are already claimed, there is an entire universe of (according to present knowledge) unclaimed, unowned resources in space.

Inheritance Day is a promising proposal for distribution

Generally in the past, unoccupied land has become owned via homesteading, in which the prospective owner occupies the land physically and develops it. However, in principle homesteading destroys economic value by giving rise to competition to become the entity performing the homesteading, which is a deadweight loss compared to the economic benefits of making use of the resources once they are claimed. The Inheritance Day proposal described by Drexler suggests that humanity select a day on which every human being alive that day is assigned an equal share of the as-yet-unclaimed resources of the universe. According to the Coase Theorem, given clear title to resources and multilateral ownership, ignoring transaction cost problems, subsequent trade leads to a good utilization of those claimed resources and to a Pareto efficient outcome for all parties, regardless of the initial distribution of resources (Coase, 1960).

Inheritance Day could provide individuals with the necessary capital to increase the likelihood that civilization will still serve human interests once AGI is reached and their skills are no longer economically sufficient.

3.3. Implementing Inheritance Day

Drexler quote on Inheritance Day implementation

When describing Inheritance Day, Drexler states that "this involves distributing ownership of the resources of space (genuine, permanent, transferable ownership) equally among all people—but doing so only once, then letting people provide for their progeny (or others') from their own vast share of the wealth of space. This will allow different groups to pursue different futures, and it will reward the frugal rather than the profligate. It can provide the foundation for a future of unlimited diversity for the indefinite future, if active shields are used to protect people from aggression and theft" (Drexler, 1986).

Timing of release of Inheritance Day assets to individuals

One idealistic interpretation of the proposal, not recommended here, is that individuals receive full title to their entire share of newly-assigned resources with complete ability to trade immediately. The Coase Theorem seems to suggest that this would be the most economically efficient solution. However, the problem is that most individuals are not yet experienced at managing capital, much less ownership of space resources. Historically, when the wealthy plan to leave an inheritance to children who are still underage, they create a trust which gradually releases the resources to benefit the beneficiary until that beneficiary has crossed an age threshold such that the grantor is willing to entrust them with the rest of the wealth. With regard to Inheritance Day, understanding that at the moment that Inheritance Day is implemented none of the beneficiaries are as yet experienced at managing capital of this nature, it would be advisable to grant some of those resources immediately to individuals, including the ability to trade, but to hold most of the resources in trust and gradually release them over time. This would enable all individuals to continue to have a gradual stream of capital that can be invested and produce financial returns as the economy continues to grow. This would help ensure that individuals are protected from making terribly egregious, early foolish mistakes.

Defining an equal share of space resources

Defining what constitutes an equal share of the resources of space is a hard problem that remains unsolved to date. What counts as an equal share relies on individuals' assessments of that share and these subjective values can differ greatly (Harms, 1989). However, a promising protocol for division is the "I cut, you choose" principle for envy-free distribution of resources with agents which have different preferences. One agent divides the resources, the other partner chooses first, and the divider receives the remaining share. While there are some new algorithms for cake-cutting for multiple agents with multiple preferences, the distribution of all resources in space remains a complex problem (Aziz, 2016). Another role model would be the approach to the privatization of resources in Poland via national wealth management funds when transitioning from a communist economy to a market economy. The government retained some of the shares of the newly privatized enterprises, gave some to the company's employees, and distributed the rest to competing National Wealth Management Funds, so that one investment group had primary responsibility for modernizing a given enterprise. From these funds, 27 million individuals received vouchers, equivalent to American-style mutual funds (Goldman, 2016).

When speaking on AI risk, Jaan Tallinn references a thought experiment involving negotiations between humanity and a powerful alien fleet which doesn't care about humanity. He says: "Even if we could secure just one galaxy out of the 100 billions as consolation prize for the losers, this would translate into 50 personal star systems for every human alive today. This illustrates two things: (1) Even if we mostly screw up, things might turn up to be pretty okay in the end and (2) the worst we can do is continue our current political zero-sum games, which cost us 50 galaxies per second" (Tallinn, 2017). Both points hold for Inheritance Day in a similar way: (1) Even if some might deem the details such as the choice of date or exact distribution of space resources as arbitrary, the sheer size of space could still eventually allow every person to be well-off, and (2) continuing to delay action and perpetuating the current system is just as much a decision as taking action to change, and is one that is potentially costly.

Inheritance Day is orthogonal to redistribution questions

The proposal for redistribution of resources, e.g., a basic income as supported by Elon Musk, Sam Altman, and other prominent figures (Agreelist, 2017) and similar to the current experiment in Finland, is a separate issue. These two approaches to attempting to ensure human financial well-being—a one-time gradual distribution of unclaimed resources, and a continual redistribution of already-owned resources—are in principle unrelated. Either could be implemented on its own, or they could be combined. Whether either or both actually become implemented are political decisions for society to make.

Conclusions

Biotechnology risks can be seen as a subset of longer-term and more challenging Stage 4 nanotechnology risks; both derive from systems of molecular machines. Similarly, cyber risks can be seen as a subset of later AGI risks. Computer security is identified as important across many risk domains. Defensive, decentralized, bottom-up, open source approaches are suggested for addressing a variety of risk areas. Inspiration can be provided by analogies with successful defense scenarios across domains, from the immune system in biology to the U.S. Constitution in politics. Timing estimates for these anticipated powerful technologies vary widely, therefore it is advisable to attempt to find strategies that are robust across these differing time estimates. A gradual, one-time distribution of unclaimed resources could help ameliorate the concern that human labor becomes much less valuable in a world with AGI. These concepts are presented as options for consideration and possible elaboration, rather than as complete policy recommendations.

Acknowledgements

We thank James C. Bennett, Steve Burgess, Ben Goertzel, Tanya Jones, Richard Mallah, Gayle Pergamit, Glenn Reynolds, Marcia Seidler, Leif Smith, Bas Steunebrink, Roman Yampolskiy, and Eliezer Yudkowsky for helpful comments; any remaining errors are our own. We thank Foresight Institute for financial support and UCLA's B. John Garrick Institute for the Risk Sciences for organizing the First Colloquium on Existential and Catastrophic Risk which stimulated this project.

References

Aguirre, Anthony. 2017. Metaculus website,
http:www.metaculus.com

Agreelist. 2017. "Universal Basic Income - Do you agree?" *Agreelist.com*
http://www.agreelist.com/s/universal-basic-income-kbupbtnz4sek

Anderson, Michael; Anderson, Susan. 2007. *Machine Ethics*. Cambridge.

Armstrong, Stuart. 2014. *Smarter Than Us.* Machine Intelligence Research Institute.

Aziz, Haris; Mackenzie, Simon. 2016. "A Discrete and Bounded Envy-Free Cake Cutting Protocol for Any Number of Agents". Foundations of Computer Science, 2016 IEEE Annual Symposium.

Bennett, James C. 2017. Private communication to Christine Peterson.

Brin, David. 1998. *The Transparent Society.* Perseus Books.

Buterin, Vitalik. 2016. "Hard Fork Completed." Ethereum Blog.
https://blog.ethereum.org/2016/07/20/hard-fork-completed/

Bostrom, Nick. 2017. "Interactions between the AI Control Problem and the Governance Problem." Talk at the 2017 Asilomar Conference, organized by FLI.
https://www.youtube.com/watch?v=_H-uxRq2w-c

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

References

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. Vol 2, ed I.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." Journal of Evolution and Technology, Vol. 9, No 1.

Brooks, Teresa Nicole, 2017. Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems. arXiv preprint
https://arxiv.org/abs/1702.06162

Cairns-Smith, A.G., 1982. *Genetic takeover and the mineral origins of life* . Wiley.

Chui, Michael; Manyika, James; Miremadi, Mehdi. 2015. Four Fundamentals of Workplace Automation. McKinsey:
http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/four-fundamentals-ofworkplace-automation

Coase, Ronald. 1960. "The Problem of Social Cost." The Ronald Coase Institute.
http://www.coase.org/coaseinterview.htm

Dennis, Jack; Van Horn, Earl. "Programming semantics for multiprogrammed computations" Communications of the ACM, 1966.

Dowd, Maureen. 2017. "Elon Musk's Billion-Dollar Crusade to Stop the AI Apocalypse." Vanity Fair , March 26, 2017.
http://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x

Drexler, K. Eric. 1986. Engines of Creation: The Coming Era of Nanotechnology . Anchor Books Edition.
http://e-drexler.com/d/06/00/EOC/EOC_Table_of_Contents.html

Drexler, K. Eric. 1988. "Nanotechnology and Exploratory Engineering," Stanford University course taught spring quarter 1988.

Drexler, K. Eric. 2007. "The Stealth Threat," *Bulletin of the Atomic Scientists* , Jan/Feb 2007.

Drexler, K. Eric. 2015. "MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities", Technical Report #2015-3, Future of Humanity Institute.
http://www.fhi.ox.ac.uk/wp-content/uploads/MDL-Intelligence-Distillation-for-safe-superintelligent-problem-solving1.pdf

Drexler, K. Eric. 2017. "Reframing the AI safety landscape". Google doc:
https://docs.google.com/document/d/145yJBoNTYHOJ_FMOO2hO-x2KnJQT45hxhtd0I84HVLE/edit

Drexler, K. Eric; Pamlin, Dennis. 2013. Nano-solutions for the 21st Century. Low Carbon Leaders.
http://www.oxfordmartin.ox.ac.uk/publications/view/1349

References

Duettmann, Allison. 2014. "The Reflective Equilibrium as Ethical Standard for AI." Master's Thesis in the Department of Philosophy, Logic and Scientific Method, London School of Economics.

Executive Office of the President. 2016. *National Electric Grid Security and Action Plan* . https://www.whitehouse.gov/sites/whitehouse.gov/files/images/National_Electric_Grid_Action_Plan_06Dec2016.pdf

Filardo, Nathaniel W., 2016, May. Research Report: Mitigating LangSec Problems with Capabilities. In Security and Privacy Workshops (SPW), 2016 IEEE (pp. 189-197). IEEE.

Freudenberg, Drew; Tirole, Jean. 1991. *Game Theory* . MIT Press.

Fisher, Kathleen. 2014. Using formal methods to enable more secure vehicles: DARPA's HACMS program. *Proceedings of the 19th ACM SIGPLAN international conference on Functional programming*

Friend, Tad. 2016. "Sam Altman's Manifest Destiny." The New Yorker , October 10, 2016. http://www.newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny

Garrick, B. John. 2008. *Quantifying and Controlling Catastrophic Risks* . Academic Press.

Goertzel, Ben. 2014. "Artificial General Intelligence: Concept, State of the Art, and Future." *Journal of Artificial General Intelligence 5*(1) https://www.degruyter.com/downloadpdf/j/jagi.2014.5.issue-1/jagi-2014-0001/jagi-2014-0001.xml

Goertzel, Ben. 2015 "Superintelligence: Fears, Promises and Potentials." Journal of Evolution & Technology 25(2) http://jetpress.org/v25.2/goertzel.htm

Goldman, Minton. 2016. *Revolution and Change in Central and Eastern Europe: Political, Economic, and Social Challenges* . Routledge.

Goodfellow, Ian; Papernot, Nicolas; Huang, Sandy; Duan, Yan; Abbeel, Pieter; Clark, Jack. 2017. "Attacking Machine Learning With Adversarial Examples." OpenAI. https://blog.openai.com/adversarial-example-research/

Gustin, Sam. 2014. "IBM: We Haven't Given the NSA Any Client Data." *Time Magazine* . http://time.com/25410/ibm-nsa-letter/

Hanson, Robin. 2003. "Combinatorial Information Market Design." Information Systems Frontiers 5(1).

Hanson, Robin. 2014. "When the Economy Transcends Humanity." World Future Society. http://mason.gmu.edu/~rhanson/Futurist.pdf

Harms, Tracy. 1989. "Dividing the Cosmic Pie." Nanocon Proceedings, Appendices. http://www.nanoindustries.com/nanojbl/NanoConProc/nanocon12.html

References

Harris, Sam. 2016. Surviving the Cosmos. Sam Harris Blog:
https://www.samharris.org/blog/item/surviving-the-cosmos

Kornwitz, Jason. 2017. "The cybersecurity risk of self-driving cars." Phys.org, Feb. 16, 2017.
https://phys.org/news/2017-02-cybersecurity-self-driving-cars.html

Kosal, Margaret. 2009. *Nanotechnology for Chemical and Biological Defense* . Springer

Kurzweil, Ray. 2014. "Don't Fear Artificial Intelligence." *Time.*
http://time.com/3641921/dont-fear-artificial-intelligence/

Kurzweil, Ray. 2012. *How To Create A Mind* . Penguin Books

Levy, Stephen. 2015 "How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over." Backchannel , Dec. 11, 2015.
https://backchannel.com/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-takingover-17e0e27dd02a

Li, Fei-Fei. 2016. "Why AI Needs Diversity". *IEEE.*
http://spectrum.ieee.org/tech-talk/at-work/tech-careers/computer-vision-leader-feifei-li-on-why-ai-needs-diversity

Mallah, Richard. 2017. "The Landscape of AI Safety and Beneficence Research. Input for Brainstorming at Beneficial AI 2017." *Beneficial AI 2017.*
https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf

McLarty. Thomas F; Ridge, Thomas L. 2014. *Securing the U.S. Electrical Grid* . Center for the Study of the Presidency & Congress.
https://www.thepresidency.org/sites/default/files/Final%20Grid%20Report_0.pdf

Merkle, Ralph. 1995. "Design-ahead". *Proceedings of the first general conference of Nanotechnology* . John Wiley and Sons.

Meselson, Matthew S.; Ratner, Daniel; Ratner, Mark A.; Drexler, K. Eric. 2015. "Emerging Technologies." *Bulletin of the Atomic Scientists* .
http://www.tandfonline.com/doi/abs/10.1080/00963402.2007.11461048

Miller, Mark S.; Drexler, K. Eric. 1988. "Comparative Ecology: A Computational Perspective". *The Ecology of Computation* . Elsevier Science Publishers

Miller, Mark S. 1997. "Computer Security as the Future of Law". Extro 3 Conference.
https://drive.google.com/file/d/0Bw0VXJKBgYPMS0J2VGIyWWlocms/edit

Miller, Mark S.; Yee, Ka-Ping; Shapiro, J. 2003. "Capability Myths Demolished". Technical Report SRL2003-2, Systems Research Laboratory, Johns Hopkins University.
http://srl.cs.jhu.edu/pubs/SRL2003-02.pdf

References

Miller, Mark S.; Van Cutsem, Tom; and Tulloh, Bill, 2013, Distributed electronic rights in JavaScript. In *European Symposium on Programming* (pp. 1-20). Springer Berlin Heidelberg.

Miller, Mark S.; Tulloh, Bill. 2016. "Decision Alignment: Large programs as complex organizations (keynote)." *Proceedings of ECOOP 2016.*
http://2016.ecoop.org/event/ecoop-2016-papers-plenary-speaker-

Minsky, Marvin. 1985. *The Society of Mind* . Simon & Schuster..

Muehlhauser, Luke. 2013. "When Will AI Be Created?" Machine Intelligence Research Institute.
https://intelligence.org/2013/05/15/when-will-ai-be-created/

Nordholz, J.C. and Seifert, J.P., 2011. Efficient Virtualization on Hardware with Limited Virtualization Support.
https://www.isti.tu-berlin.de/fileadmin/fg214/finished_theses/Nordholz/diplom_nordholz.pdf

Ng, Andrew. 2015. "Why Deep Learning is a Mandate for Humans, Not Just For Machines". *Wired.*
https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/

O'Cleirigh, Fiona. 2015. " Bill Binney, the 'original' NSA whistleblower, on Snowden, 9/11 and illegal surveillance." *Computer Weekly* .
http://www.computerweekly.com/feature/Interview-the-original-NSA-whistleblower

Omohundro, Steve. 2014. "Autonomous technology and the greater good." *Journal of Experimental & Theoretical Artificial Intelligence. Vol 26* .
http://www.tandfonline.com/doi/full/10.1080/0952813X.2014.895111%20

Open Philanthropy Project. 2017. "OpenAI—General Support," March 2017.
http://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/openai-general-support

OpenAI. 2017. "About Open AI"
https://openai.com/about/#mission

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined* . Penguin Group.

Popper, Karl R. 1945. *The Open Society and its Enemies.* Routledge.

Rashid, Fahmida Y. 2015. " Cyber Attack on Power Grid Could Top $1 Trillion in Damage: Report." *Security Week* , July 16, 2015.
http://www.securityweek.com/cyber-attack-power-grid-could-top-1-trillion-damage-report

Risley, James. 2015. "Elon Musk, Peter Thiel, Reid Hoffman and others commit $1B to stop AI from taking over the world." *Geekwire* , Dec. 11, 2015.

Russell, Stuart; Norvig, Peter. 2009. *Artificial Intelligence: A Modern Approach.* Pearson Education Limited.

References

Simpson, Corbin; Short Allen. "Monte Language", https://github.com/monte-language/monte 2017.

Soares, Nate. 2016. "The Value Learning Problem." *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence* . https://intelligence.org/files/ValueLearningProblem.pdf

Steunebrink, Bas; Thorisson, Kristinn; Schmidhuber, Juergen. 2016. "Growing Recursive Self-Improvers." *AGI Conference 2016* . http://people.idsia.ch/~steunebrink/Publications/AGI16_growing_recursive_self-improvers.pdf

Szigeti, Balasz., et. al. 2014. " OpenWorm: an open-science approach to modeling Caenorhabditis elegans". *Front. Comput. Neurosc.*, Nov. 3, 2014.

Tallinn, Jaan. 2012. "CSaP Distinguished Lecture: The intelligence stairway." *University of Cambridge* . http://www.csap.cam.ac.uk/events/csap-distinguished-lecture-intelligence-stairway/

Tallinn, Jaan. 2017. "AI and Value Alignment." 2017 *Beneficial AI Conference*. https://www.youtube.com/watch?v=d6plk-JxfGw

Urban, Tim. 2015. "The AI Revolution: The Road to Superintelligence." WaitButWhy. http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

Watson, Robert N., Anderson, J., Laurie, Ben. and Kennaway, K., 2012a. A taste of Capsicum: practical capabilities for UNIX. *Communications of the ACM* , 55(3), pp. 97-104.

Watson, Robert N., Neumann, Peter G., Woodruff, J., Anderson, J., Anderson, Ross, Dave, N., Laurie, Ben, Moore, S.W., Murdoch, S.J., Paeps, P. and Roe, M., 2012b, March. CHERI: a research platform deconflating hardware virtualization and protection. In *Workshop paper, Runtime Environments, Systems, Layering and Virtualized Environments (RESoLVE 2012)* .

Weber, Robert. 2014. Open Letter to Clients. IBM Blog https://www.ibm.com/blogs/think/2014/03/open-letter-data/

Wikileaks, 2013. "Nations Buying As Hackers Sell Computer Flaws." https://wikileaks.org/hackingteam/emails/emailid/96008

Yampolskiy, Roman; Spellchecker, M. 2016. " Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures." https://arxiv.org/abs/1610.07997v1

Yee, Ka-Ping. 2007. "Building Reliable Voting Machine Software." *A dissertation submitted to the Graduate Division of the University of California, Berkeley* . http://zesty.ca/pubs/yee-phd.pdf

Yudkowsky, Eliezer. 2004. "Coherent Extrapolated Volition." Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf

References

Yudkowsky, Eliezer. 2015a. *Rationality: From AI to Zombies*. Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2015b. "The Value Loading Problem." Edge.
https://www.edge.org/response-detail/26198

Yudkowsky, Eliezer. 2016a. "Corporations vs. Superintelligences." Arbital.
https://arbital.com/p/corps_vs_si/

Yudkowsky, Eliezer. 2016b. " Task-directed AI". Arbital.
https://arbital.com/p/task_agi/

FORESIGHT INSTITUTE

Cyber, Nano, and AGI Risks: Decentralized Approaches to Reducing Risks

by Christine Peterson, Mark S. Miller, and Allison Duettmann