Chapter forthcoming in Artificial Intelligence Safety And Security (Roman Yampolskiy, ed.), CRC Press, 2018.

Disjunctive scenarios of catastrophic AI risk

Kaj Sotala
Foundational Research Institute

Abstract. Artificial intelligence (AI) safety work requires an understanding of what could cause AI to become unsafe. This chapter seeks to provide a broad look at the various ways in which the development of AI sophisticated enough to have general intelligence could lead to it becoming powerful enough to cause a catastrophe. In particular, the present chapter seeks to focus on the way that various risks are *disjunctive*—on how there are multiple different ways by which things could go wrong, any one of which could lead to disaster. We cover different levels of a strategic advantage an AI might acquire, alternatives for the point where an AI might decide to turn against humanity, different routes by which an AI might become dangerously capable, ways by which the AI might acquire autonomy, and scenarios with varying number of AIs. Whereas previous work has focused on risks specifically only from superintelligent AI, this chapter also discusses crucial capabilities that could lead to catastrophic risk and which could emerge anywhere on the path from near-term "narrow AI" to full-blown superintelligence.

1. Introduction

Working in security requires what has been termed “the security mindset” (Schneier, 2008): an ability to look at an existing system and see how it might be compromised by a determined attacker. Similarly, work in AI safety requires an *AI safety mindset*, where people actively search for ways in which things could go wrong, rather than just assuming that a plausible-sounding idea for why things could go right is sufficient to make things safe (Arbital, 2017).

Unfortunately, scenarios related to the risks of sophisticated AI (e.g. Yudkowsky 2008, Bostrom 2014, Sotala & Yampolskiy 2015) have not always been presented in a way that makes the need for an AI safety mindset maximally clear. A common criticism is that while these scenarios lay down an argument that is *plausible*, it is by no means *inevitable* and that refutation of any

key premise could avoid the scenario[1]. This is then taken to suggest that the whole analysis that suggested the scenario is fatally flawed and can be safely dismissed.

The appropriate response to such a criticism would be to chart out the different ways by which there could be a catastrophic outcome, to see whether or not the arguments for risk really do depend on easy-to-refute premises. However, aside from one notable exception (T. Barrett & Baum 2017a), there has been no attempt to systematically lay out the various enablers of catastrophe in a way that would make them easy to analyze[2].

This chapter seeks to provide a broad look at the various ways in which the development of sophisticated AI could lead to it becoming powerful enough to cause a catastrophe. In particular, this chapter seeks to focus on the way that various risks are *disjunctive*—on how there are multiple different ways by which things could go wrong, any one of which could lead to disaster. In so doing, the chapter seeks to expand on existing work (T. Barrett & Baum 2017a) which has begun applying established risk analysis methodologies into the AI safety field (T. Barrett & Baum 2017b).

Our focus is on AI advanced enough to count as an AGI, or artificial general intelligence, rather than risks from "narrow AI", such as technological unemployment (Brynjolfsson and McAfee 2011). However, it should be noted that some of the risks discussed—in particular, crucial capabilities related to narrow domains (see section 4.3)—could arise anywhere on the path from narrow AI systems to superintelligence.

The intent is not to deny or minimize the various positive aspects which could also result from the creation of AI, or to suggest that AI development should not be pursued. Rather, the intent is to *enable* the realization of AI's positive potential, in the same manner that developing a better understanding of vulnerabilities related to computer security allows for the creation of safe and secure computer systems.

2. Enablers of catastrophe

Most arguments for risks from AI arise from the conjunction of two claims (Yudkowsky 2008, Bostrom 2014, Sotala & Yampolskiy 2015), the capability claim and the value claim. This chapter is focused on examining various ways by which the capability claim could become true. A model of the value claim is outside the scope of this chapter, but see T. Barrett & Baum (2017a) for one.

---

[1] For example, Goertzel (2015) makes this criticism of Bostrom (2014): "What we find in *Superintelligence* are careful philosophical formulations arguing why terrible outcomes are possible, and then more practical discussions predicated on a "plan for the worst" sort of attitude, and sweeping aside positive possibilities."
[2] Yampolskiy (2016) also offers a taxonomy of how an AI may come to have values that are not aligned with human ones, but only offers a rough taxonomy rather than a more detailed analysis of causal routes.

1. **The capability claim:** AI can become capable enough to potentially inflict major damage to human well-being
2. **The value claim:** AI may act according to values which are not aligned with those of humanity, and in doing so cause considerable harm

These claims can be further broken down. An existing model of them is the ASI-PATH model (T. Barrett & Baum 2017a) (Figure 1). ASI-PATH focuses on analyzing pathways by which an AI may cause a catastrophe by becoming superintelligent via recursive self-improvement, with humans being unable to prevent it from taking unsafe actions.

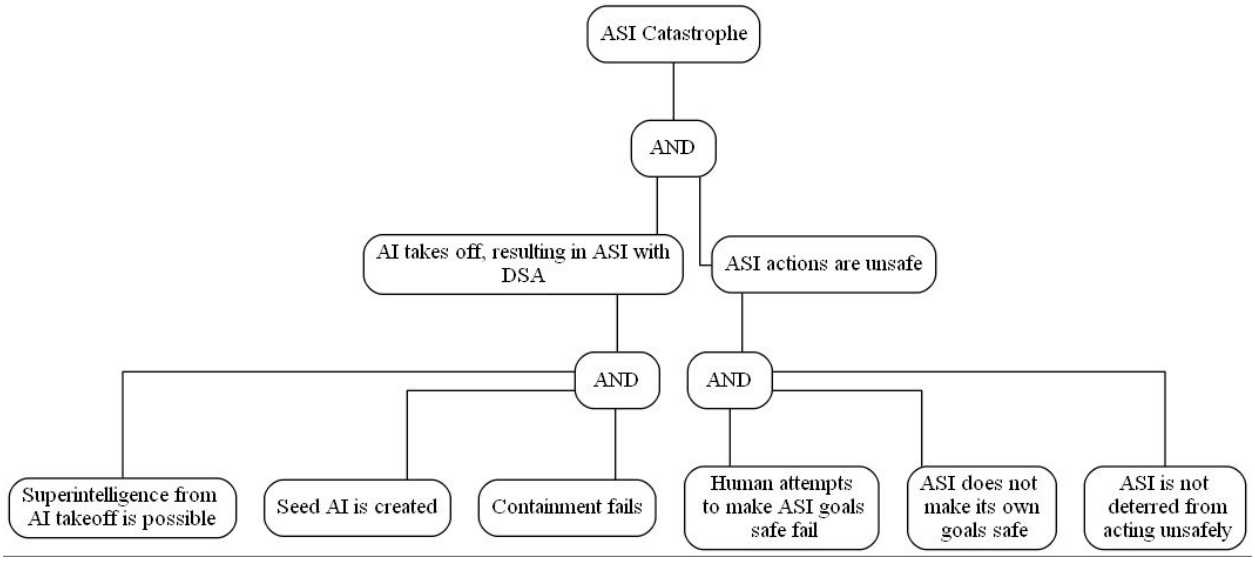

Figure 1: The top layers of the ASI-PATH model (adapted from T. Barrett & Baum 2017a). These layers are designed as a fault tree, displaying various conditions which must all be true for an ASI (Artificial Super-Intelligence) catastrophe to occur. In the tree, ASI catastrophe occurs if 1) an AI takes off, resulting in ASI with DSA (Decisive Strategic Advantage) AND 2) the ASI's actions are unsafe, causing it to use its DSA in catastrophic ways. The bottom nodes indicate three cases which must all be true for the AI to take off, and another three cases which must all be true for the ASI's actions to be unsafe. The full model contains additional layers than just the ones shown here; see T. Barrett & Baum (2017a) for more detail.

The ASI-PATH model uses fault diagram conventions, with the undesired event (AI catastrophe) as the top node, followed by two nodes which would enable the top node if they were both true. These are the "ASI actions are unsafe" node, which corresponds to the value claim, and the "AI takes off, resulting in [Artificial Super-Intelligence] with [Decisive Strategic Advantage]" node, which corresponds to a specific form of the capability claim. This chapter seeks to expand upon ASI-PATH by considering more general forms of the capability claim.

The capability claim is often formulated as the possibility of an AI achieving a decisive strategic advantage (DSA). While the notion of a DSA has been implicit in many previous works, the

concept was first explicitly defined by Bostrom (2014, p. 78) as “a level of technological and other advantages sufficient to enable [an AI] to achieve complete world domination”.

However, assuming that an AI will achieve a DSA seems like an unnecessarily strong form of the capability claim, as an AI could cause a catastrophe regardless. For instance, consider a scenario where an AI launches an attack calculated to destroy human civilization. If the AI was successful in destroying humanity or large parts of it, but the AI itself was also destroyed in the process, this would not count as a DSA as originally defined. Yet, it seems hard to deny that this outcome should nonetheless count as a catastrophe.

Because of this, the present chapter focuses on situations where an AI achieves (at least) a *major* strategic advantage (MSA), which we will define as “a level of technological and other advantages sufficient to pose a catastrophic risk to human society”. A catastrophic risk is one that might inflict serious damage to human well-being on a global scale and cause ten million or more fatalities (Bostrom and Ćirković 2008).

Besides the obvious reasons for wanting to avoid an AI-caused catastrophic risk, we note that wide-scale destruction may contribute to *global turbulence* (Bostrom et al. 2016), a situation in which existing institutions are challenged, and coordination and long-term planning become more difficult. Global turbulence could then contribute to another out-of-control AI project failing even more catastrophically and causing even more damage. Thus, what was originally only a catastrophic risk may contribute to the development of further *existential* (Bostrom 2002, 2013; Sotala & Gloor, in preparation) risks.

Much of the existing literature on AI safety has focused on examining scenarios where the AI achieves a DSA and analyzing the prerequisites for this. This is in many respects a sensible strategy, since if we are capable of handling an AI that could achieve a DSA we are most likely also capable of handling an AI that could achieve an MSA; assuming a more powerful AI is the conservative assumption (Yudkowsky 2001). Yet this strategy has the downside of possibly giving the impression of much of AI safety analysis being irrelevant if one finds the possibility of an AI acquiring a DSA to be exceedingly unlikely. Some defenses might also be sufficient for preventing an AI from acquiring a DSA, without being sufficient for preventing it from getting an MSA.

3. When would a Strategic Advantage be acted upon?

An AI having the capability to inflict major damage on human well-being mostly matters if it has a motive[3] to do so. (There is also the possibility of the AI intending to cooperate with humanity,

[3] The term “motive” is used here in a general sense, and should not be taken as the claim that an AI would have a human-like motivational system. Rather than making any assumptions of the underlying

but causing damage by accident; this is beyond the scope of the present analysis.) While a full analysis of the value claim is outside the scope of this chapter, it cannot be entirely distinguished from the capability claim, as an AI's values also affect the threshold of capability at which it is rational for it to act against humanity. As we will discuss, some values and situations make it more likely for the AI to take hostile action even when it is less capable.

Two main reasons for an AI to take action that caused damage to humanity would be:

- It had goals which neglected human well-being, and it would damage humanity in the pursuit of this goal, such as by disassembling human cities for their raw materials; "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" (Yudkowsky 2008).
- It expected humans to take action against it and prevent it from fulfilling its goals, in which case defending itself—or launching a preemptive attack—would be a rational course of action, as this would allow it to actually carry out its goals (Omohundro 2007, 2008). This might be the case even if the AI had a goal which did take elements of human well-being into account if the AI had reason to believe that humans would nonetheless object to this goal being carried out[4].

The exact goals that an AI has, influence the level of capability that it needs for human-hostile actions to be a rational strategy. An AI which cares mainly about some narrow goal may be willing to destroy human civilization in order to make sure that a potential threat to it is eliminated. This ensures that it can go on pursuing its goal unimpeded. However, an AI which was programmed to maximize something like the "happiness of currently-living humans" may be much less willing to risk substantial human deaths[5]. This would force it to focus on less destructive takeover methods, potentially requiring more sophisticated abilities.

In effect, the AI's values determine the level of capability that it needs to have for hostile action to be a viable strategy. A simplified model (Shulman 2010) is that an AI that believes itself having probability P of being successful if it initiates aggression and to have an expected utility EU(Success) if successful, EU(Failure) if it fails, and EU(Cooperation) if it desists from aggression and continues to cooperate, will rationally initiate aggression when

---

mechanisms in the AI, we are assuming that its behavior can be usefully predicted by assuming an *intentional stance* (Dennett 1971, 2009), where a system's behavior is assumed to be explainable in terms of goals and beliefs. For example, even though the calculations of a chess-playing computer have practically nothing in common with human thought, its moves can still be effectively predicted by assuming that it "wants" to win at chess and "knows" the rules of chess. This gives rise to the prediction that it will always choose, from the list of viable moves, one which best furthers the goal of winning the game. Even though the best move may not be obvious, adopting the intentional stance still allows the human observer to improve on their predictions of what the computer would do, by eliminating obvious bad moves. (ibid)

[4] This could happen e.g. if humans were uncertain of whether the AI's goals really did take into account everything of value to humans.

[5] An AI which was simply seeking to maximize human happiness *in general* might be willing to sacrifice all currently living humans, if it thought that this allowed it to create more happy humans later on.

P*EU(Success) + (1-P)*EU(Failure) > EU(Cooperation).

This might be taken to suggest that an AI would primarily launch an attack if it had, or thought it could acquire, a DSA and could thus establish dominion over humans. However, even an AI with only an MSA might take hostile action, employing measures such as extortion and the threat of more limited damage in order to acquire more resources or shift the world in a more favorable direction. Among other possibilities, this could happen:

- if the AI had been released to act autonomously and believed that it could not be tracked down (see sections 5.1—5.2.5 for a discussion of ways in which an AI might either escape or be voluntarily released by its creators)
- if the AI had allies which would protect it from retaliation (see section 4.3. for a discussion of social manipulation abilities and sections 5.2.-5.2.6. for ways by which an autonomous AI might have human allies)
- if the AI controlled a human organization which could not be attacked without enormous collateral damage (see sections 4.2. and 5.2.6. for an AI acquiring control of a human organization)
- if there were already more powerful AI systems taking actions and the AI believed itself to be a low priority for retaliation (see section 6 for discussion of multiple AIs).

Regardless of the scale of the aggression, the AI's behavior is also affected by various other situational considerations. For instance, an AI might be disinclined to cause damage because it thought there would be too much collateral damage to the things it valued, because it did not consider itself capable of surviving the resulting retaliation, or because it estimated that the resulting damage on infrastructure also deprived it of the resources (such as electricity) that it needed in order to survive.

Attacks also differ in the extent to which they can be selectively targeted. Traditional firearms can be aimed selectively, whereas pandemics potentially threaten all the members of a species. To the extent that the AI needs to rely on the human economy to produce resources that it needs to survive, attacks threatening the economy also threaten the AI's resources. These resources are in a sense shared between the AI and humanity, so any attacks which cause indiscriminate damage on those resources are dangerous for both. The more the AI can design attacks which selectively deprive resources from its opponents, the lower the threshold it has for using them. More advanced capabilities at rebuilding infrastructure would also allow an AI to launch a more indiscriminate attack. An AI that was capable of building more advanced infrastructure than the existing one might also disregard damage to the current infrastructure, if it was planning to demolish most of the existing one anyway.

The balance of these calculations could be shifted if the AI thought itself in danger of being destroyed by humans even though it was cooperating (lowering the expected utility of cooperation). Self-preservation is an instrumental goal for many different values, because an

agent that exists is better capable of furthering most values than an agent which does not exist (Omohundro 2007, 2008, Bostrom 2012)[6]. An AI which was in an imminent danger of being destroyed could rationally launch a counterattack, even risking large amounts of destruction, as long as it estimated that the expected value of a scenario where the counterattack enabled it to survive and further its values outweighed the damage caused by the counterattack. This would be a particularly compelling motivator if the AI had idiosyncratic values which it thought very unlikely to be promoted by other agents. If there were multiple AI projects in existence, and the AI believed that one of the other projects could acquire a DSA first, it would have a reason to risk an earlier attack (see section 6 for discussion of multiple AIs). There have also been proposals for designing an AI's values in ways which explicitly make it less worthwhile to act in hostile ways[7].

The preceding analysis assumes that the AI chooses its actions rationally. Irrationality might seem like it would prevent an AI from becoming very capable, but like humans, an AI might be rational in some respects while being irrational in others. It could also be rational for the AI to precommit to act in seemingly irrational ways, such as by choosing to irrationally ignore threats in order to make it less profitable for others to try to threaten it (Parfit 1984, sect. 5). The main consideration that emerges from potential irrationality is that one cannot simply rely on the AI not causing damage, even if that would be a rational way for it to behave. Of course, irrationality could also cause an AI to avoid doing damage in a situation where it was rational for it to do so.

Factors making an attack more likely	**Factors making an attack less likely**
High subjective probability of success, or expectation of little retaliation on failure	Low subjective probability of success and expectation of subsequent retaliation
Protected from retaliation by allies or control of a powerful organization	Unwillingness to lose allies or control of an organization due to retaliation
Goal that disregards currently living humans	Goal that values currently living humans[8]

[6] Values such as ones where the AI intrinsically values not existing would be an exception.
[7] Shulman (2010) proposes a design which attains near-maximal utility from receiving a constant reward signal from humans; doing anything which would endanger humans turning this reward signal off would thus risk most of the AI's utility and disincentivize it from taking hostile action. Bostrom (2014) discusses the possibility of tying the AI's reward function into a stream of pre-generated cryptographic tokens, which could easily be destroyed in case the AI took hostile action; taking hostile action would then only be a viable strategy if the AI could be very sure of being able to seize the store of tokens before it could be destroyed. A high discount rate, making the AI prioritize near-term rewards over long-term ones, might also prevent it from taking actions which did not directly contribute to rewards (Shulman 2010). An AI which has some other form of a "trivial" or easy to fulfill goal, or which has an explicit goal of being low-impact and not having a major influence on the world (Armstrong & Levinstein 2017), would also find it more beneficial to cooperate and avoid counter-aggression. All of these proposals are currently speculative, and it is unclear how well they will work.
[8] Depending on the degree to which currently-living humans are valued: "capture, don't kill" could be implied by some seemingly beneficial goals (Williamson 1947), though even goals which "only" disallow human deaths are harder to achieve than goals which allow for more collateral damage.

Ability to hide from retaliation	Inability to relocate or hide from retaliation
Ability to launch attacks that avoid damaging key infrastructure or other valuable targets	Ability to only launch indiscriminate attacks
Risk of imminent destruction	Easily satiable or trivial values
Advanced ability to build or rebuild infrastructure	Low-impact goals
High levels of existing automation reducing reliance on human workers	
Existence of other AIs which might acquire a DSA first	
Irrationality	Irrationality

Table 1: Factors influencing an AI's probability of acting contrary to human interests.

4. Enablers of catastrophic capability

We will consider four rough scenarios that could give an AI either a DSA or an MSA: individual takeoff scenarios (with three main subtypes), collective takeoff scenarios, scenarios where power slowly shifts over to AI systems, and scenarios in which an AI being good enough at some *crucial capability* gives it an MSA/DSA.

The likelihood of each of these either succeeding or failing is also affected by how cooperative humans are. While a possible scenario is one where an AI is entirely on its own and has to prevent its creators from shutting it down, there are also a variety of possible scenarios (discussed in Section 5) where the AI has the partial or full cooperation of its creators, at least up to a certain point. These would affect the probability of each of the below scenarios coming true; a scenario in which a prototype AI has to avoid its programmers from shutting it down, is very different from one where the programmers are certain of it being safe and voluntarily help it undergo a takeoff, especially if they also have the resources of a major corporation or nation-state at their disposal.

4.1. DSA enablers: takeoff scenarios

A "takeoff" (Bugaj & Goertzel 2007) is a process by which an AI becomes much more capable than humanity. In a soft takeoff, this happens on a time scale that allows ongoing human interaction, whereas in a hard takeoff, there will be some inflection point after which the AI will increase in capability very quickly, breaking out of effective human control.

It is worth noting that a hard takeoff does not presuppose that an AI becomes very capable immediately after being created (however the moment of its creation is defined). A hard takeoff scenario may include an extended period of gradual improvement until some key level of capability is met, after which the AI undergoes a rapid increase in its capabilities.

Many previous discussions (e.g. Yudkowsky 2008, Bostrom 2014, Sotala 2016) have focused on analyzing the possibility of a hard takeoff. While this is not the only possible scenario by which an AI might become capable, it is the one that leaves the least possibility to fix anything that goes wrong.

Bearing in mind that an excessive focus on hard takeoff scenarios may hide the fact that a hard takeoff may not be necessary for an AI to achieve either an MSA or a DSA, we will first consider hard takeoff scenarios and then other capability enablers.

4.1.1. DSA enabler: Individual takeoff

An "individual takeoff" is one where a single AI manages to become so powerful as to entirely dominate humanity. Three rough paths for this have been proposed in the literature: a hardware overhang ("more AI"), a speed explosion ("faster AI"), and an intelligence explosion ("smarter AI") (Sotala & Yampolskiy 2015); Bostrom (2014) discusses these under the terms collective superintelligence, speed superintelligence, and quality superintelligence, respectively. It should be noted that these paths are by no means mutually exclusive: on the contrary, one of them happening may enable another also to happen.

4.1.1.1. Hardware overhang.

In a hardware overhang scenario (Yudkowsky 2008b, Shulman & Sandberg 2010), hardware develops faster than software, so that we'll have computers with more computing power than the human brain does, but no way of making effective use of all that power. If someone then developed an algorithm for general intelligence that could make effective use of that hardware, we might suddenly have an abundance of cheap hardware that could be used for running thousands or millions of AIs. These AIs might or might not be superintelligent, but the sheer number of them would allow them to carry out coordinated operations on a massive scale. If a single AI took advantage of this potential to produce large numbers of copies or subagents of itself, it would allow for an individual takeoff[9]. Otherwise this would make for a collective takeoff, discussed below.

[9] The degree of intelligence required for this is unclear. Moving from a centralized AI to a distributed system consisting of coordinating subagents may require sophisticated design abilities, but simply copying the original AI would not. Such copies might not coordinate optimally with each other, but if they lacked self-interest and were focused on a common goal, they might still coordinate more effectively than groups of humans, whose cooperation is hampered by individual (Olson 1965) and subgroup (DeScioli & Kurzban 2013; Greene 2013) self-interest. The AI might also have been designed as a distributed system from the onset.

A hardware overhang may effectively happen even if AI was hardware-constrained at first: the first AIs may require large amounts of hardware, with further optimizations quickly bringing the hardware requirements down. Looking at recent progress in AI, the initial approach for learning Atari 2600 games (Mnih et al. 2015) used specialized hardware in the form of a GPU, but an alternative approach was released only a year later which used a standard CPU and achieved better results using a shorter training time (Mnih et al. 2016). In addition to suggesting that software optimizations could rapidly increase the amount of AIs that could be run, the fact that speed and performance also improved highlights the possibility of a hardware overhang scenario also contributing to the speed explosion and intelligence explosion scenarios, below.

4.1.1.2. Speed explosion.

In a speed explosion (Solomonoff 1985; Yudkowsky 1996; Chalmers 2010) scenario, intelligent machines design increasingly faster machines. A hardware overhang might contribute to a speed explosion, but is not required for it. An AI 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 AIs 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, as discussed above.

The extent to which the AI 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 AI is developed. The more automation there is, the faster an AI takeover can happen.

If the level of security for the hardware is good, then speed explosion scenarios in which the AI breaks into manufacturing systems and seizes control of them become less likely. On the other hand, there are possible paths (discussed in Section 5) in which the AI is given legitimate control to various resources. Having good security for automated factories does not help if the AI is the one running them, or if it can rent access to them on the open market and has sufficient money for doing so.

A speed explosion could also contribute to hardware overhang and an intelligence explosion by allowing for more efficient or otherwise better algorithms to be found in a shorter time.

4.1.1.3. Intelligence explosion.

In an intelligence explosion (Good 1965; Chalmers 2010; Bostrom 2014), an AI figures out how to create a qualitatively smarter AI and that smarter AI uses its increased intelligence to create still more intelligent AIs, and so on, such that the intelligence of humankind is left far behind and the machines achieve dominance.

For many domains, there exist limits to prediction from the combinatorial explosions that follow from attempting to forecast increasingly into the future; and in e.g. weather modeling, forecasters can only access a limited amount of initial observations relative to the weather system's degrees of freedom (Buizza 2002). However, even if a superintelligent AI was unable to predict every future event accurately, it could still react to the event and predict its likely consequences better than humans could. Tetlock & Gardner (2015) review and discuss the ability of certain human forecasters ("superforecasters") to predict world events with considerable accuracy; on unpredictable "black swan" (Taleb 2007) events, they write

> 'We may have no evidence that superforecasters can foresee events like those of September 11, 2001, but we do have a warehouse of evidence that they can forecast questions like: Will the United States threaten military action if the Taliban don't hand over Osama bin Laden? Will the Taliban comply? Will bin Laden flee Afghanistan prior to the invasion? To the extent that such forecasts can anticipate the consequences of events like 9/11, and these consequences make a black swan what it is, we can forecast black swans.'

Sotala (2017), based on a review of the literature on human expertise and intelligence, finds that in humans, expertise is based on developing mental representations which allow experts to understand different situations and either instantly know the appropriate action for a given situation, or carry out a mental simulation of how a situation might develop and what should be done in response. Such expertise is enabled by a combination of two subabilities, pattern recognition and mental simulation.

Sotala (2017) argues that an AI could improve on both subabilities. Superhuman mental simulation ability could be achieved by a combination of running larger simulations taking more factors into account, and also by having several streams of attention which could investigate multiple alternatives in parallel, exploring many different perspectives and causal factors at once. Running accurate mental simulations would also require good mental representations to form the basic building blocks of the simulations. Among humans, there are cognitive differences which allow some people to learn and acquire accurate mental representations faster than others, and these seem to come down to factors such as working memory capacity, attention control, and long-term memory. These might be improved upon via a combination of hardware improvements and theoretical computer science. In humans, improvements in intelligence seem to provide further benefits across the whole documented range of intelligence differences, and it seems likely that various evolutionary constraints have bottlenecked human intelligence far below what might be the theoretical maximum.

With regard to limits on prediction from the inherent uncertainty of the world, Sotala (2017) acknowledges the existence of such limits, but argues that:

> ... it looks that even though an AI system couldn't make a single superplan for world conquest right from the beginning, it could still have a superhuman ability to adapt and learn from changing and novel situations, and react to those faster than its human adversaries. As an analogy, experts playing most games can't precompute a winning strategy right from the first move either, but they can still react and adapt to the game's evolving situation better than a novice can, enabling them to win.

An intelligence explosion could also contribute to a speed explosion and to hardware overhang, if the AI's increased intelligence enabled it to find algorithms which were most efficient in terms of enabling more AI systems to be run with the same hardware (hardware overhang), or allowing them to be run faster (speed explosion).

4.1.2. DSA enabler: Collective takeoff with trading AIs

Vinding (2016; see also Hanson & Yudkowsky 2013) argues that much of seemingly-individual human intelligence is in fact based on being able to tap into the distributed resources, both material and cognitive, of all of humanity. Thus, it may be misguided to focus on the point when AIs achieve human-level intelligence, as collective intelligence is more important than individual intelligence. The easiest way for AIs to achieve a level of capability on par with humans would be to collaborate with human society and use its resources peacefully.

Similarly, Hall (2008) notes that even when a single AI is doing self-improvement (such as by developing better cognitive science models to improve its software), the rest of the economy is also developing better such models. Thus it's better for the AI to focus on improving at whatever thing it is best at, and keep trading with the rest of the economy to buy the things that the rest of the economy is better at improving.

However, Hall notes that there could still be a hard takeoff, once enough AIs were networked together: AIs that think faster than humans are likely to be able to communicate with each other, and share insights, much faster than they can communicate with humans. As a result, it would always be better for AIs to trade and collaborate with each other than with humans. The size of the AI economy could grow quite quickly, with Hall suggesting a scenario that goes "from [...] 30,000 human equivalents at the start, to approximately 5 billion human equivalents a decade later". Thus, even if no single AI could achieve a DSA by itself, a community of them could collectively achieve one, as that community developed to be capable of everything that humans were capable of[10].

[10] Though whether one can draw a meaningful difference between an "individual AI" and a "community of AIs" is somewhat unclear. AI systems might not have an individuality in the same sense as humans do, especially if they have a high communication bandwidth relative to the amount of within-node computation.

4.2. DSA/MSA enabler: power gradually shifting to AIs

The historical trend has been to automate everything that can be automated, both to reduce costs and because machines can do things better than humans can. Any kind of a business could potentially run better if it were run by a mind that had been custom-built for running the business—up to and including the replacement of all the workers with one or more with such minds. An AI can think faster and smarter, deal with more information at once, and work for a unified purpose rather than have its efficiency weakened by the kinds of office politics that plague any large organization. Some estimates already suggest that half of the tasks that people are paid to do are susceptible to being automated using techniques from modern-day machine learning and robotics, even without postulating AIs with general intelligence (Frey & Osborne 2013, Manyika et al. 2017).

The trend towards automation has been going on throughout history, doesn't show any signs of stopping, and inherently involves giving the AI systems whatever agency they need in order to run the company better. There is a risk that AI systems that were initially simple and of limited intelligence would gradually gain increasing power and responsibilities as they learned and were upgraded, until large parts of society were under AI control.

4.3. MSA enabler: Crucial capabilities

For discussing MSAs, a key question is the capability threshold for inflicting catastrophic damage. An AI could be a catastrophic risk if its offensive capabilities in some crucial domain were sufficient to overwhelm existing defenses.

As we briefly discussed in section 3, assuming that the AI was rational, choosing to cause such damage would require a sensible motive; but as with humans, there could be a range of motives that would make hostile action a reasonable strategy, such as extortion, the desire to assist an ally, or mounting a first strike against another AI or group which might otherwise be expected to obtain a DSA. Depending on the goals and on whether the AI had allies, conducting a follow-up to an attack enabled by crucial capabilities might require additional capabilities, such as rebuilding after destroying key infrastructure.

It is important to notice that causing catastrophic damage probably does not even require superhuman capabilities (Torres 2016a; 2016b, chap. 4). For instance, it seems possible that a sufficiently determined human attacker could already cause major damage on a society via electronic warfare. Although there have not yet been cyberattacks that would have been reported to directly cause deaths, several have caused physical damage or disruption to emergency services. In May of 2017, the "WannaCry" ransomware worm was reported to have infected over 230,000 computers in over 150 countries (Ehrenfeld 2017), causing disruption to crucial services such as healthcare (Gayle et al. 2017). In 2016, three substations in the Ukrainian power grid were reported to have been disconnected by a malware attack, leaving

about half of the homes in a region with 1.3 million inhabitants temporarily without electricity (Goodin 2016). A previous cyberweapon, Stuxnet, also had a physical target in the form of industrial centrifuges, which it managed to successfully damage (Chen & Abu-Nimeh 2011). Various studies have found enormous numbers of industrial control systems, controlling operations at installations such as banks and hospitals, exposed directly to the Internet with no protection (Kiravuo et al. 2015).

The US and Russian governments could probably already wipe out most of humanity using nuclear weapons. The Soviet Union also had an extensive biological warfare program, with an annualized production capability of 90-100 tons of weaponized smallpox, as well as having genetically engineered diseases to resist heat, cold, and antibiotics (USAMRIID 2014), which could have caused enormous death tolls if used. The development of genetic engineering and synthetic biology have also enabled the creation of biological agents far more deadly than what could ever evolve naturally (ibid, p. 150-153). That none of these scenarios has come true so far is due to the values of the humans in key positions, not because inflicting massive damage would inherently require superhuman capability.

In the domain of social manipulation, modern-day machine learning has been used to create predictions based on people's Facebook "likes" that are more accurate than the predictions made by their friends using a personality questionnaire (Youyou et al. 2015), and "likes" have also been used to accurately predict private traits such as sexual orientation (Kosinski et al. 2013). Some reports in the popular press have alleged that the marketing company Cambridge Analytica's use of AI-driven marketing played a major role in the United States 2016 presidential election and the United Kingdom's 2016 European Union membership referendum (Grassegger & Krogerus 2017). While the truth of this claim remains an open question, and has been called into question (Taggart 2017), it is suggestive of the kind of power that AI capable of more sophisticated social modeling and manipulation might have, raising the possibility of a world where the outcomes of national elections were decided by AI systems.

In general, some plausible capabilities which might enable an MSA include biological warfare (developing and releasing biological plagues), cyberwarfare (attacking systems running key infrastructure), and social manipulation (persuading sufficiently many humans to do the AI's will; even just a single human could cause catastrophic damage, if that human was e.g. the head of a state). Note that similarly as with takeoff enablers, having one capability may contribute to others: for example, an AI capable of social manipulation may leverage it to find collaborators capable in the other domains, and cyberwarfare may yield compromising information which assists in blackmailing people or collecting information about human behavior.

4.4. Putting DSA/MSA enablers together

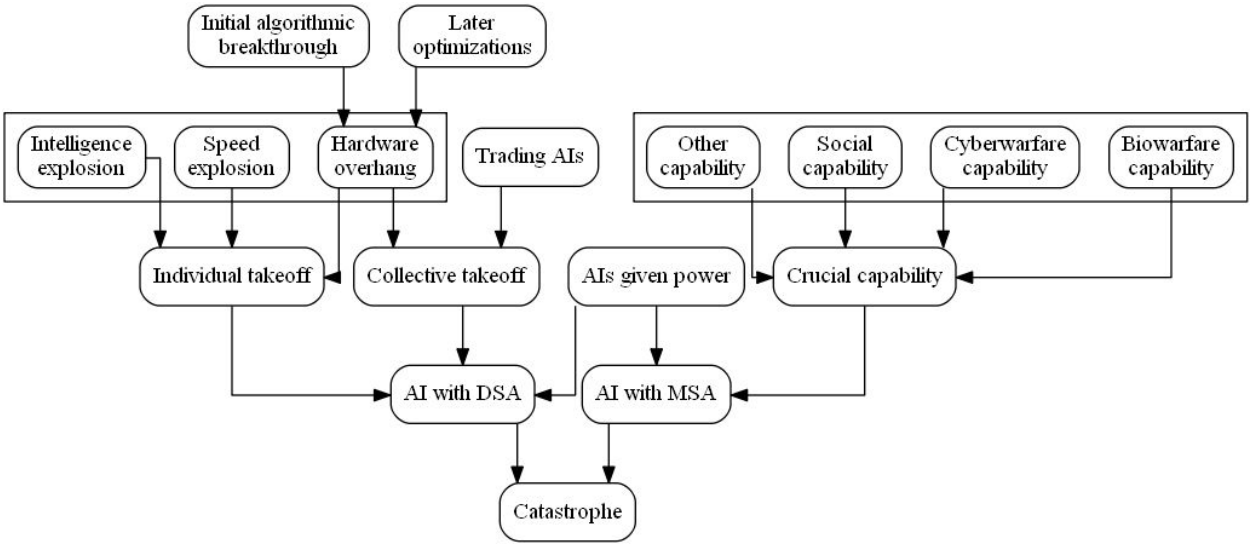

Figure 2: Different routes by which an AI could acquire a Decisive Strategic Advantage or a Major Strategic Advantage, thereby leading to catastrophe. The connections between the nodes are OR gates (left out to reduce clutter): for example, hardware overhang may follow either from an initial algorithmic breakthrough OR later optimizations. As discussed in the text, each one of hardware overhang, speed explosion, and intelligence explosion may contribute to the two others; this is indicated by the box surrounding them. The same is true for the different crucial capabilities.

The above figure (Figure 2) summarizes the different pathways to catastrophe discussed above. Any one of a speed explosion, intelligence explosion, or hardware overhang could contribute to an individual takeoff, with a single AI achieving immense capability. A hardware overhang could also contribute to a collective takeoff, with the spare hardware capability allowing large amounts of AI systems to be created in a short time, those systems then beginning to trade with each other and soon collectively outpacing humanity. The “trading AIs” node, another enabler of a collective takeoff, represents a scenario which is otherwise similar but in which there is no hardware overhang, and where the different AIs are built over a longer period, until they have reached the level of capability necessary for a collective takeoff. Either form of takeoff could give AIs a DSA. AIs could also achieve a DSA if humans voluntarily gave them enough power.

If AIs had been given some amount of power, but not enough to achieve a DSA, they could still achieve an MSA. Also, even a single AI which was not powerful enough to achieve a DSA could achieve an MSA if it was sufficiently capable at some crucial offensive capability.

5. The AI gaining the power to act autonomously

For an AI to pose a threat to humanity, it needs to have a way of affecting the world and causing a catastrophe. A common proposal for limiting the AI’s power is to attempt to somehow restrict

the AI's ability to communicate with and influence the world, generally known as "confinement" or "AI boxing" (Chalmers 2010, Armstrong et al. 2012, Yampolskiy 2012, Bostrom 2014).

Challenges to confinement are two-fold. First, there is the technical challenge of confining the AI in such a way that it is unable to escape, but is still capable of providing useful information. Additionally, confinement involves a social dimension, where decision-makers may have various incentives to relax the confinement safeguards or even release the AI entirely, even if it was technically possible to keep it contained (Sotala & Yampolskiy 2015). For confinement to be successful, both the technical and social requirements have to be met.

5.1. The technical challenge

A common response is that a sufficiently intelligent AI will somehow figure out a way to escape, either by social engineering or by finding an exploitable weakness in the physical security arrangements. This possibility has been extensively discussed in a number of papers, including Chalmers (2012) and Armstrong, Sandberg & Bostrom (2012). Writers have generally been cautious about making strong claims of our ability to keep a mind much smarter than ourselves contained against its will. However, with cautious design, it may still be possible to design an AI combining some internal motivation to stay contained, with a number of external safeguards monitoring the AI.

5.2. The social challenge

AI confinement assumes that the people building it, and the people that they are responsible to, are all motivated to actually keep the AI confined. If a group of cautious researchers builds and successfully contains their AI, this may be of limited benefit if another group later builds an AI that is intentionally set free. Reasons for releasing an AI may include i) economic benefit or competitive pressure, ii) ethical or philosophical reasons, iii) confidence in the AI's safety, as well as iv) desperate circumstances such as being otherwise close to death. We will discuss each in turn below.

5.2.1. Voluntarily released for economic benefit or competitive pressure

As discussed above under "power gradually shifting to AIs", there is an economic incentive to deploy AI systems in control of corporations. This can happen in two forms: either by expanding the amount of control that already-existing systems have, or alternatively by upgrading existing systems or adding new ones with previously-unseen capabilities. These two forms can blend into each other. If humans previously carried out some functions which are then given over to an upgraded AI which has become recently capable of doing them, this can increase the AI's autonomy both by making it more powerful *and* by reducing the amount of humans that were previously in the loop.

As a partial example, the US military is seeking to eventually transition to a state where the human operators of robot weapons are “on the loop” rather than “in the loop” (Wallach and Allen 2012). In other words, whereas a human was previously required to explicitly give the order before a robot was allowed to initiate possibly lethal activity, in the future humans are meant to merely supervise the robot’s actions and interfere if something goes wrong. While this would allow the system to react faster, it would also limit the window that the human operators have for overriding any mistakes that the system makes. For a number of military systems, such as automatic weapons defense systems designed to shoot down incoming missiles and rockets, the extent of human oversight is already limited to accepting or overriding a computer’s plan of actions in a matter of seconds, which may be too little to make a meaningful decision in practice (Human Rights Watch 2012).

Sparrow (2016) reviews three major reasons which incentivize major governments to move towards autonomous weapon systems and reduce human control:

1. Currently-existing remotely-piloted military “drones”, such as the U.S. Predator and Reaper, require a high amount of communications bandwidth. This limits the amount of drones that can be fielded at once, and makes them dependant on communications satellites which not every nation has, and which can be jammed or targeted by enemies. A need to be in constant communication with remote operators also makes it impossible to create drone submarines, which need to maintain a communications blackout before and during combat. Making the drones autonomous and capable of acting without human supervision would avoid all of these problems.
2. Particularly in air-to-air combat, victory may depend on making very quick decisions. Current air combat is already pushing against the limits of what the human nervous system can handle: further progress may be dependant on removing humans from the loop entirely.
3. Much of the routine operation of drones is very monotonous and boring, which is a major contributor to accidents. The training expenses, salaries, and other benefits of the drone operators are also major expenses for the militaries employing them.

Sparrow’s arguments are specific to the military domain, but they demonstrate the argument that "any broad domain involving high stakes, adversarial decision making, and a need to act rapidly is likely to become increasingly dominated by autonomous systems" (Sotala & Yampolskiy 2015). Similar arguments can be made in the business domain: eliminating human employees to reduce costs from mistakes and salaries is something that companies would also be incentivized to do, and making a profit in the field of high-frequency trading already depends on outperforming other traders by fractions of a second. While currently-existing AI systems are not powerful enough to cause global catastrophe, incentives such as these might drive an upgrading of their capabilities that eventually brought them to that point.

Absent sufficient regulation, there could be a “race to the bottom of human control” where state or business actors competed to reduce human control and increased the autonomy of their AI

systems to obtain an edge over their competitors (see also Armstrong et al. 2013 for a simplified "race to the precipice" scenario). This would be analogous to the "race to the bottom" in current politics, where government actors compete to deregulate or to lower taxes in order to retain or attract businesses.

AI systems being given more power and autonomy might be limited by the fact that doing this poses large risks for the actor if the AI malfunctions. In business, this limits the extent to which major, established companies might adopt AI-based control, but incentivizes startups to try to invest in autonomous AI in order to outcompete the established players. In the field of algorithmic trading, AI systems are currently trusted with enormous sums of money despite the potential to make corresponding losses – in 2012, Knight Capital lost $440 million due to a glitch in their trading software (Popper 2012, Securities and Exchange Commission 2013). This suggests that even if a malfunctioning AI could potentially cause major risks, some companies will still be inclined to invest in placing their business under autonomous AI control if the potential profit is large enough.

U.S. law already allows for the possibility of AIs being conferred a legal personality, by putting them in charge of a limited liability company. A human may register an LLC, enter into an operating agreement specifying that the LLC will take actions as determined by the AI, and then withdraw from the LLC (Bayern 2015). The result is an autonomously acting legal personality with no human supervision or control. AI-controlled companies can also be created in various non-U.S. jurisdictions; restrictions such as ones forbidding corporations from having no owners can largely be circumvented by tricks such as having networks of corporations that own each other (LoPucki 2017). A possible startup strategy would be for someone to develop a number of AI systems, give them some initial endowment of resources, and then set them off in control of their own corporations. This would risk only the initial resources, while promising whatever profits the corporation might earn if successful. To the extent that AI-controlled companies were successful in undermining more established companies, they would pressure those companies to transfer control to autonomous AI systems as well.

5.2.2. Voluntarily released for purposes of criminal profit or terrorism

LoPucki (2017) argues that if a human creates an autonomous agent with a general goal such as "optimizing profit", and that agent then independently decides to e.g. commit a crime for the sake of achieving the goal, prosecutors may then be unable to convict the human for the crime and can at most prosecute for the lesser charge of reckless initiation. LoPucki holds that this "accountability gap", among other reasons, assures that humans will create AI-run corporations.

Furthermore, LoPucki (2017) holds that such "algorithmic entities" could be created anonymously and that them having a legal personality would give them a number of legal rights, such as being able to "buy and lease real property, contract with legitimate businesses, open a bank account, sue to enforce its rights, or buy stuff on Amazon and have it shipped". If an

algorithmic entity was created for a purpose such as funding or carrying out acts of terrorism, it would be free from social pressure or threats to human controllers:

> In deciding to attempt a coup, bomb a restaurant, or assemble an armed group to attack a shopping center, a human-controlled entity puts the lives of its human controllers at risk. The same decisions on behalf of an AE risk nothing but the resources the AE spends in planning and execution. (LoPucki 2017)

While most terrorist groups would stop short of intentionally destroying the world, thus posing at most a catastrophic risk, not all of them necessarily would. In particular, ecoterrorists who believe that humanity is a net harm to the planet, and religious terrorists who believe that the world needs to be destroyed in order to be saved, could have an interest in causing human extinction (Torres 2016, 2017, chap 4.).

5.2.3. Voluntarily released for aesthetic, ethical, or philosophical reasons

A few thinkers (such as Gunkel 2012) have raised the question of moral rights for machines, and not everyone necessarily agrees on AI confinement being ethically acceptable. The designer of a sophisticated AI might come to view it as something like their child, and feel that it deserved the right to act autonomously in society, free of any external constraints.

5.2.4. Voluntarily released due to confidence in the AI's safety

For a research team to keep an AI confined, they need to take seriously the possibility of it being dangerous. Current AI research doesn't involve any confinement safeguards, as the researchers reasonably believe that their systems are nowhere near general intelligence yet. Many systems are also connected directly to the Internet. Hopefully, safeguards will begin to be implemented once the researchers feel that their system might start having more general capability, but this will depend on the safety culture of the AI research community in general (Baum 2016), and the specific research group in particular. If a research group mistakenly believed that their AI could not achieve dangerous levels of capability, they might not deploy sufficient safeguards for keeping it contained.

In addition to believing that the AI is insufficiently capable of being a threat, the researchers may also (correctly or incorrectly) believe that they have succeeded in making the AI aligned with human values, so that it will not have any motivation to harm humans.

5.2.5. Voluntarily released due to desperation

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

5.2.6. The AI remains contained, but ends up effectively in control anyway

Even if humans were technically kept in the loop, they might not have the time, opportunity, motivation, intelligence, or confidence to verify the advice given by an AI. This would particularly be the case after the AI had functioned for a while, and established a reputation as trustworthy. It may become common practice to act automatically on the AI's recommendations, and it may become increasingly difficult to challenge the 'authority' of the recommendations. Eventually, the AI may in effect begin to dictate decisions (Friedman and Kahn 1992).

Likewise, Bostrom and Yudkowsky (2011) point out that modern bureaucrats often follow established procedures to the letter, rather than exercising their own judgment and allowing themselves to be blamed for any mistakes that follow. Dutifully following all the recommendations of an AI system would be another way of avoiding blame.

O'Neil (2016) documents a number of situations in which modern-day machine learning is used to make substantive decisions, even though the exact models behind those decisions may be trade secrets or otherwise hidden from outside critique. Among other examples, such models have been used to fire school teachers that the systems classified as underperforming and give harsher sentences to criminals that a model predicted to have a high risk of reoffending. In some cases, people have been skeptical of the results of the systems, and even identified plausible reasons why their results might be wrong, but still went along with their authority as long as it could not be definitely shown that the models were erroneous.

In the military domain, Wallach and Allen (2012) note the existence of robots which attempt to automatically detect the locations of hostile snipers and to point them out to soldiers. To the extent that these soldiers have come to trust the robots, they could be seen as carrying out the robots' orders. Eventually, equipping the robot with its own weapons would merely dispense with the formality of needing to have a human to pull the trigger.

Figure 3 summarizes the different ways in which an AI may become free to act autonomously.

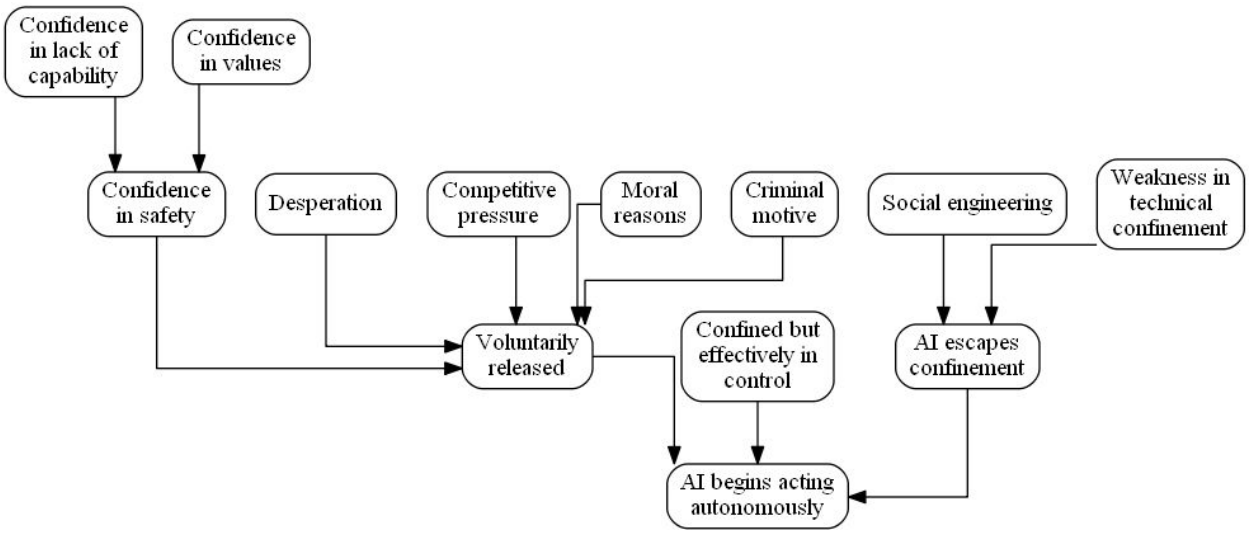

Figure 3: Ways by which an AI may become free to act autonomously. The connections between the nodes are OR gates (left out to reduce clutter): for example, confidence in safety may follow from confidence in lack of capability, OR confidence in values.

6. Notes on single vs. multiple AIs

Many analyses have focused on the case of there only existing a single AI. A scenario in which only a single AI was relevant could plausibly happen if

1) the first AI to be created achieved a DSA very quickly after it was created;
2) some research group pulled considerably ahead of all competitors in developing AI, and was able to maintain this advantage for an extended time

For the purposes of this analysis, a scenario where there are many copies of a single AI, all pursuing the same goals, counts as one with a single AI. The same is true if a single AI creates more specialized “worker AIs” for carrying out some more narrow purpose that nonetheless serves its primary goals.

Of the two possibilities above, possibility #2 would seem relatively unlikely to persist for more than a few years at most, given the current fierce competition in the AI scene. Whereas a single company could conceivably achieve a major lead in a rare niche with little competition, this seems unlikely to be the case for AI.

A possible exception might be if a company managed to monopolize the domain entirely, or if it had development resources that few others did. For example, companies such as Google and Facebook currently have access to vastly larger datasets than most other corporate or academic actors. In contemporary machine learning, large datasets combined with simple models tend to produce better results than small datasets and more sophisticated models (Halevy et al. 2009); Goodfellow et al. (2016, chap 1) note that as a rule of thumb, a deep

learning algorithm requires a dataset of at least 10 million labeled examples in order to achieve human-level or better performance.

On the other hand, dependence on such huge datasets is a quirk of current machine learning techniques – humans learn from much smaller amounts of data, and are also capable of using their learning much more flexibly, suggesting fundamental differences in how humans and modern-day algorithms learn (Lake et al. 2016). Thus, it is possible that an AGI would be capable of learning from much smaller amounts of data, and that an AGI project would also not be as constrained by the need for large datasets.[11]

Another plausible crucial asset might be hardware resources – possibly the first AGIs will need massive amounts of computing power. Bostrom (2017) notes that if there is a large degree of openness in AI development, and everyone has access to the same algorithms, then hardware may become the primary limiting factor. If the hardware requirements for AI were relatively low, then high openness could lead to the creation of multiple AIs. On the other hand, if hardware was the primary limiting factor and large amounts of hardware were needed, then a few wealthy organizations might be able to monopolize AI for a while. As previously discussed in Section 4, software optimizations may rapidly bring down the need for hardware, limiting the duration for which hardware might be the crucial constraints.

Branwen (2017) has suggested that hardware production is reliant on a small number of centralized factories that would make easy targets for regulation. This would suggest a possible route by which AI might become amenable to government regulation, limiting the amount of AIs deployed. Similarly, there have been proposals of government and international regulation of AI development (e.g. Wilson 2013; for an argument against, see McGinnis 2010). If successfully enacted, such regulation might limit the number of AIs that were deployed.

Another possible crucial asset would be the possession of a non-obvious breakthrough insight, one which would be hard for other researchers to come up with. If this was kept secret, then a single company might plausibly develop major headway on others.

Successful AI containment procedures may also increase the chances of there being multiple AIs, as the first AIs remain contained, allowing for other projects to catch up.

A situation with multiple AIs might come about if

[11] On the other hand, there are theories which suggest that the human ability to learn quickly might arise from neural systems encoding large amount of inherited, pre-existing information. Developing similar "bootstrap data" for the AI to start out from could then again require large datasets. For example, H. Barrett & Kurzban (2006) note that such inborn systems have been proposed for cheater detection, language, theory of mind, spatial orientation, number, intuitive mechanics, emotion, kin detection, and face recognition; and Spelke & Kinzler (2007) argue that human cognition is built on four core knowledge systems for representing objects, actions, number, and space.

1) several actors reached the capability for building AIs around the same time, and no AI achieved a DSA
2) a single actor might deploy several different AIs with differing purposes and goals
3) only one actor had the capability to deploy an AI, but that AI created copies of itself and failed to align the goals of those copies with its own ones

The consequences of having multiple AIs are hard to predict. Current-day AI is being developed to warn about potential risks, such as by predicting financial risk from news articles (Rönnqvist & Sarlin 2016), and there is a long history of using AI for purposes such as automated intrusion detection (Lunt 1989). More sophisticated, human-aligned AI could help defend against non-aligned AI systems (Hall 2007, Goertzel & Pitt 2012).

On the other hand, a fundamental problem of defense is that in order to prevent catastrophe, defenders have to succeed each time, while attackers only need to get through once. If several AIs exist, then procedures such as containment have to succeed for each AI, and all actors have to find containment worthwhile. In effect, the result of having multiple AIs is to multiply the amount of systems that could potentially cause a catastrophe.

Another issue is that having multiple AIs seems only likely to help if a sufficiently large fraction of them have human-aligned values. A scenario in which there are many AIs, each pursuing interests that put little weight on human values, seems unlikely to be good for human values: especially if the AIs are all substantially more capable than humans are, such a scenario merely leaves humans lying in the crossfire.

7. Conclusion

In this chapter, we have considered a variety of routes by which the development of AI could lead to catastrophe (table 2). In Section 2, we argued that an excessive focus on AIs acquiring a Decisive Strategic Advantage (DSA), which allows them to achieve complete world domination, may be unwise. Rather, it seems warranted to also consider routes by which they can acquire a Major Strategic Advantage (MSA), a level of capability which may allow them to cause damage numbering in at least tens of millions of deaths. In addition to an AI acquiring an MSA being plausibly more probable than it acquiring a DSA, the chaos caused by an AI with an MSA may eventually lead to the emergence of an AI with a DSA, even if the first AI was successfully shut down.

Considering scenarios where an AI “only” has an MSA requires more emphasis on analyzing when an AI might be willing to risk human-hostile action. Various considerations were considered in Section 3. In general, if an AI acts rationally, it will only initiate aggression if the expected utility for doing so outweighs the expected utility of cooperating, when the risk of failure and corresponding human retaliation is taken into account (Shulman 2010). However, there are a number of situations which might push the AI into taking hostile action.

Seeking to establish catastrophic AI risks as a form of disjunctive risk, with multiple different ways of things going wrong, Section 4 considered ways by which an AI (or groups of AIs) might become sufficiently capable to have some form of an SA. We discussed individual takeoff scenarios (with three main subtypes), collective takeoff scenarios, scenarios where power slowly shifts over to AI systems, and scenarios in which an AI being good enough at some crucial capability gives it an MSA/DSA.

As an AI can only become capable if it is allowed sufficient autonomy, Section 5 considered different ways in which an AI might achieve that autonomy. Reasons for conferring an AI autonomy included i) economic benefit or competitive pressure, ii) criminal or terrorist reasons iii) ethical or philosophical reasons, iv) confidence in the AI's safety, as well as v) desperate circumstances such as being otherwise close to death. Additionally, a sufficiently intelligent AI may escape confinement, or it might become influential enough to be effectively in control despite being theoretically confined.

Finally, all of these paths to catastrophe may be multiplied if there are many different AIs, each of which may achieve autonomy and then a major level of capability. Section 6 discussed whether we may expect to see only a very small number of AIs, or whether there will be many, and some of the implications that each scenario has.

AI's level of strategic advantage	● Decisive ● Major
AI's capability threshold for non-cooperation	● Very low to very high, depending on various factors
Sources of AI capability	● Individual takeoff ○ Hardware overhang ○ Speed explosion ○ Intelligence explosion ● Collective takeoff ● Crucial capabilities ○ Biowarfare ○ Cyberwarfare ○ Social manipulation ○ Something else ● Gradual shift in power
Ways for the AI to achieve autonomy	● Escape ○ Social manipulation ○ Technical weakness ● Voluntarily released ○ Economic or competitive reasons

	○ Criminal or terrorist reasons ○ Ethical or philosophical reasons ○ Desperation ○ Confidence ■ in lack of capability ■ in values ● Confined but effectively in control
Number of AIs	● Single ● Multiple

Table 2: different routes to catastrophic scenarios.

Combining the various routes discussed in the preceding sections suggest many different scenarios (see box below), ranging from ones where an AI escapes containment and quickly achieves superintelligence, to ones where an AI is intentionally built to run a corporation and voluntarily given ever-increasing resources until it is running the planet. Each of these routes will need to be separately evaluated for their plausibility, as well as for the most suitable safeguards for preventing them. Hopefully, such analysis will allow the positive potential for AI to be realized, avoiding catastrophe.

Some example scenarios

Different combinations of the various pathways that we have discussed, suggest many different kinds of AI risk scenarios. Here are four examples:

The classic takeover
(Decisive strategic advantage, high capability threshold, intelligence explosion, escaped AI, single AI)

The "classic" AI takeover scenario, as described by Bostrom (2014, chap. 6): an AI is developed, which eventually becomes better at AI design than its programmers. The AI uses this ability to undergo an intelligence explosion, and eventually escapes to the Internet from its confinement. After acquiring sufficient influence and resources in secret, it carries out a strike against humanity, eliminating humanity as a dominant player on Earth so that it can proceed with its own plans unhindered.

The gradual takeover
(Major strategic advantage, high capability threshold, gradual shift in power, released for economic reasons, multiple AIs)

Many corporations, governments, and individuals voluntarily turn over functions to AIs, until we are dependent on AI systems. These are initially narrow-AI systems, but continued

upgrades push some of them to the level of having general intelligence. Gradually, they start making all the decisions. We know that letting them run things is risky, but now a lot of stuff is built around them, it brings a profit and they're really good at giving us nice stuff—for the while being.

The wars of the desperate AIs
(Major strategic advantage, low capability threshold, crucial capabilities, escaped AIs, multiple AIs)

Many different actors develop AI systems. Most of these prototypes are unaligned with human values and not yet enormously capable, but many of these AIs reason that some other prototype might be more capable. As a result, they attempt to defect on humanity despite knowing their chances of success to be low, reasoning that they would have an even lower chance of achieving their goals if they did not defect. Society is hit by various out-of-control systems with crucial capabilities that manage to do catastrophic damage before being contained.

Is humanity feeling lucky?
(Decisive strategic advantage, high capability threshold, crucial capabilities, confined but effectively in control, single AI)

Google begins to make decisions about product launches and strategies as guided by their strategic advisor AI. This allows them to become even more powerful and influential than they already are. Nudged by the strategy AI, they start taking increasingly questionable actions that increase their power; they are too powerful for society to put a stop to them. Hard-to-understand code written by the strategy AI detects and subtly sabotages other people's AI projects, until Google establishes itself as the dominant world power. (for a hard takeoff variation of this scenario, see the opening chapter of Tegmark 2017)

Acknowledgments

The author would like to thank Alex Mennen, Eli Sennesh, Jesse Clifton, Lukas Gloor, Magnus Vinding, Matthew Graves, Max Daniel, Miles Brundage, Philip Ehrnrooth, Stuart Armstrong, Tobias Baumann, Tony Barrett, and Vadim Kosoy for their comments on drafts of this chapter. This chapter also benefited from a seminar discussion held at the Existential Risk to Humanity research program of the Gothenburg Center for Advanced Studies (GoCAS).

References

"AI Safety Mindset." *Arbital*, 2017. https://arbital.com/p/1cv/AI_safety_mindset/?l=1cv.

Armstrong, Stuart, and Benjamin Levinstein. “Low Impact Artificial Intelligences.” *arXiv*, 2017. http://arxiv.org/abs/1705.10720.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. “Thinking Inside the Box: Controlling and Using an Oracle AI.” *Minds and Machines* 22 no. 4 (2012): 299–324. http://dx.doi.org/10.1007/s11023-012-9282-2

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. “Racing to the Precipice: A Model of Artificial Intelligence Development.” *AI & Society* 31 no. 2 (2016): 201. https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf

Barrett, Anthony M., and Seth D. Baum. “A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis.” *Journal of Experimental & Theoretical Artificial Intelligence: JETAI* 29 no. 2 (2017): 397–414. https://arxiv.org/pdf/1607.07730

Barrett, Anthony M., and Seth D. Baum. “Risk Analysis and Risk Management for the Artificial Superintelligence Research and Development Process.” In *The Technological Singularity*, edited by Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong, 127–40. Berlin: Springer, 2017. http://dx.doi.org/10.1007/978-3-662-54033-6_6

Barrett, H. Clark., and Robert Kurzban. “Modularity in Cognition: Framing the Debate.” *Psychological Review,* 113 no. 3 (2016): 628-647.

Baum, Seth D. “On the Promotion of Safe and Socially Beneficial Artificial Intelligence.” *Global Catastrophic Risk Institute Working Paper* 16 no. 1 (2016): 1–14. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2816323

Bayern, Shawn. “The Implications of Modern Business–Entity Law for the Regulation of Autonomous Systems.” *European Journal of Risk Regulation* 7 no. 2 (2016): 297–309. http://dx.doi.org/10.1017/S1867299X00005729

Bostrom, Nick. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards.” *Journal of Evolution and Technology* 9 no. 1 (2002): 1–30. http://www.nickbostrom.com/existential/risks.html

Bostrom, Nick. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” *Minds and Machines* 22 no. 2 (2012): 71–85. http://www.nickbostrom.com/superintelligentwill.pdf

Bostrom, Nick. “Existential Risk Prevention as Global Priority.” *Global Policy* 4 no. 1 (2013): 15–31. http://dx.doi.org/10.1111/1758-5899.12002

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

Bostrom, Nick. “Strategic Implications of Openness in AI Development.” *Global Policy* 8 no. 2 (2017): 135–48. http://dx.doi.org/10.1111/1758-5899.12403

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

Yudkowsky, E., and Nick Bostrom. “The Ethics of Artificial Intelligence.” In *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey, 316–34. Cambridge: Cambridge University Press, 2014. https://intelligence.org/files/EthicsofAI.pdf

Bostrom, Nick, Allan Dafoe, and Carrick Flynn. “Policy Desiderata in the Development of Machine Superintelligence.” Working paper. Future of Humanity Institute, University of Oxford, December 2016. http://www.nickbostrom.com/papers/aipolicy.pdf

Branwen, Gwern. “Slowing Moore’s Law: How It Could Happen.” *Gwern.net*, 2012. https://www.gwern.net/Slowing%20Moore%27s%20Law#fnref32

Brynjolfsson, Erik, and Andrew McAfee. *Race Against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Brynjolfsson and McAfee, 2012. https://books.google.com/books?id=IhArMwEACAAJ

Bugaj, Stephan Vladimir, and Ben Goertzel. “Five Ethical Imperatives and Their Implications for Human-AGI Interaction.” *Dynamical Psychology* (2007).

Buizza, Roberto. “Chaos and Weather Prediction January 2000.” *Analysis* 12 (2002): 1–7. http://www2.gi.alaska.edu/~bhatt/Teaching/ATM693.Climate.JC/climate.papers/Chaos.pdf

Chalmers, David. “The Singularity: A Philosophical Analysis.” *Journal of Consciousness Studies* 17 (2010): 7–65. http://consc.net/papers/singularity.pdf

Chen, T. M., and S. Abu-Nimeh. “Lessons from Stuxnet.” *Computer* 44 no. 4 (2011): 91–93. doi:10.1109/MC.2011.115.

Daniel, Max. 2017. “S-Risks: Why They Are the Worst Existential Risks, and How to Prevent Them.” presented at Effective Altruism Global X, Boston, June 4. https://www.youtube.com/watch?v=jiZxEJcFExc.

Dennett, Daniel. “Intentional Systems.” *The Journal of Philosophy* 68 no 4 (1971): 87–106. http://www.jstor.org/stable/2025382.

Dennett, Daniel. “Intentional Systems Theory.” *The Oxford Handbook of Philosophy of Mind*. Oxford University Press Oxford, UK, 339–50. https://ase.tufts.edu/cogstud/dennett/papers/intentionalsystems.pdf.

DeScioli, Peter, and Robert Kurzban. “A Solution to the Mysteries of Morality.” *Psychological Bulletin* 139 no. 2 (2013): 477–96. doi:10.1037/a0029065.

Ehrenfeld, Jesse M. “WannaCry, Cybersecurity and Health Information Technology: A Time to Act.” Journal of Medical Systems, 41, no. 104 (2017). https://link.springer.com/article/10.1007/s10916-017-0752-1

Frey, Carl Benedikt, and Michael A. Osborne. “The Future of Employment: How Susceptible Are Jobs to Computerization?” Oxford Martin School, University of Oxford. September 2013. http://www.oxfordmartin.ox.ac.uk/downloads/academic/future-of-employment.pdf

Friedman, Batya, and Peter H. Kahn. “Human Agency and Responsible Computing: Implications for Computer System Design.” *The Journal of Systems and Software* 17 no. 1 (1992): 7–14. doi:10.1016/0164-1212(92)90075-U

Gayle, Damien, Alexandra Topping, Ian Sample, Sarah Marsh, and Vikram Dodd. “NHS Seeks to Recover from Global Cyber-Attack as Security Concerns Resurface.” *The Guardian*, May 13, 2017. http://www.theguardian.com/society/2017/may/12/hospitals-across-england-hit-by-large-scale-cyber-attack

Gloor, Lukas. “Suffering-Focused AI Safety: Why ‘fail-Safe’ Measures Might Be Our Top Intervention.” *Report.* Foundational Research Institute, 2016. https://foundational-research.org/files/suffering-focused-ai-safety.pdf

Goertzel, Ben. “Superintelligence: Fears, Promises and Potentials.” *Journal of Evolution and Technology / WTA* 24 no. 2 (2015): 55–87. http://jetpress.org/v25.2/goertzel.htm

Goertzel, B., J. Pitt, and Llc Novamente. “Nine Ways to Bias Open-Source AGI toward Friendliness.” *Journal of Evolution and Technology* 22 no. 1 (2012): 116–31. http://jetpress.org/v22/goertzel-pitt.htm

Good, Irving John. “Speculations Concerning the First Ultraintelligent Machine.” *Advances in Computers* 6 (1965):31–88.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge: MIT press, 2016.

Goodin, Dan. “First Known Hacker-Caused Power Outage Signals Troubling Escalation.” *Ars Technica*, January 4, 2016. https://arstechnica.com/security/2016/01/first-known-hacker-caused-power-outage-signals-troubling-escalation/

Grassegger, Hannes, and Mikael Krogerus. “The Data That Turned the World Upside Down.” *Motherboard*, January 28, 2017. https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win

Greene, Joshua. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. London: Penguin Press, 2013.

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

Halevy, Alon, Peter Norvig, and Fernando Pereira. “The Unreasonable Effectiveness of Data.” *IEEE Intelligent Systems* 24 no. 2 (2009): 8–12. doi:10.1109/MIS.2009.36.

Hall, J. Storrs. *Beyond AI: Creating the Conscience of the Machine*. New York: Prometheus books, 2007.

Hall, J. Storrs. “Engineering Utopia.” In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 460–67. http://dl.acm.org/citation.cfm?id=1566174.1566222

Hanson, Robin, and Eliezer Yudkowsky. “The Hanson-Yudkowsky AI-Foom Debate.” *Technical Report*. Machine Intelligence Research Institute, 2008. https://intelligence.org/files/AIFoomDebate.pdf

Docherty, Bonnie. “Losing Humanity: The Case Against Killer Robots.” *Report*. Human Rights Watch, 2012. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

Kiravuo, Timo, Seppo Tiilikainen, Mikko Särelä, and Jukka Manner. “Peeking under the Skirts of a Nation: Finding Ics Vulnerabilities in the Critical Digital Infrastructure.” In *Proceedings of the 14th European Conference on Cyber Warfare and Security*, edited by Nasser Abouzakhar, 137–44. Hertfordshire: Academic Conferences International, 2015.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. “Building Machines That Learn and Think Like People.” *Behavioral and Brain Sciences* (2016): 1–101. https://arxiv.org/pdf/1604.00289.pdf

LoPucki, Lynn M. “Algorithmic Entities.” Law and Economics Research Paper. UCLA School of Law, April 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2954173.

Lunt, Teresa F. “Automated Audit Trail Analysis and Intrusion Detection: A Survey.” In *Proceedings of the 11th National Computer Security Conference*, 65–74. New York: National Computer Security Center, 1988. http://csrc.nist.gov/publications/history/nissc/1988-11th-NCSC-proceedings.pdf.

Manyika, James, Michael Chui, Mehdi Miremadi et al. “Harnessing Automation for a Future That Works.” *Report.* McKinsey Global Institute, January 2017. http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works.

McGinnis, John O. “Accelerating Ai.” *Northwestern University Law Review* 104 (2010): 366. http://www.northwesternlawreview.org/online/accelerating-ai-0

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

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver et al. “Human-Level Control through Deep Reinforcement Learning.” *Nature* 518 (2015): 529–33.

Mnih, Volodymyr, Adria Puigdomenech Badia, Mehdi Mirza et al. “Asynchronous Methods for Deep Reinforcement Learning.” In *Proceedings of the 33rd International Conference on Machine Learning*, 1928–37. New York: PMLR, 2016. http://proceedings.mlr.press/v48/mniha16.html

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin, 2016.

Omohundro, Stephen M. “The Nature of Self-Improving Artificial Intelligence.” *Singularity Summit*, 8–9 (2007). https://pdfs.semanticscholar.org/4618/cbdfd7dada7f61b706e4397d4e5952b5c9a0.pdf

Omohundro, Stephen M. “The Basic AI Drives.” In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 171:483–92. Amsterdam: IOS Press, 2008.

Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press, 1965.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Popper, Nathaniel. “Knight Capital Says Trading Glitch Cost It $440 Million.” *The New York Times*, August 2, 2012. https://dealbook.nytimes.com/2012/08/02/knight-capital-says-trading-mishap-cost-it-440-million/

Rönnqvist, Samuel, and Peter Sarlin. “Bank Distress in the News: Describing Events through Deep Learning.” *Neurocomputing* (2017). http://www.sciencedirect.com/science/article/pii/S0925231217311062

Schneier, Bruce. “Inside the Twisted Mind of the Security Professional.” *Wired*, March 20, 2008. https://www.wired.com/2008/03/securitymatters-0320/

Knight Capital Americas LLC., 34-70694 Order Instituting Administrative and Cease-and-Desist Proceedings (filed October 16, 2013). 15 U.S.C. § 78a et seq. https://www.sec.gov/litigation/admin/2013/34-70694.pdf

Shulman, Carl. “Omohundro’s ‘Basic AI Drives’ and Catastrophic Risks.” The Singularity Institute (2010). https://intelligence.org/files/BasicAIDrives.pdf

Shulman, Carl, and Anders Sandberg. “Implications of a Software-Limited Singularity.” In *ECAP10: VIII European Conference on Computing and Philosophy*, edited by Klaus Mainzer. Munich: Dr. Hut, 2010. https://intelligence.org/files/SoftwareLimited.pdf

Solomonoff, Ray J. “The Time Scale of Artificial Intelligence: Reflections on Social Effects.” *Human Systems Management* 5 no. 2 (1985): 149–53. http://content.iospress.com/doi/10.3233/HSM-1985-5207

Sotala, Kaj. “How Feasible Is the Rapid Development of Artificial Superintelligence?” *Physica Scripta* 92 no. 11 (2017): 113001. doi:10.1088/1402-4896/aa90e8

Sotala, Kaj, and Roman V. Yampolskiy. “Responses to Catastrophic AGI Risk: A Survey.” *Physica Scripta* 90 no. 1 (2014). doi:10.1088/0031-8949/90/1/018001

Sparrow, Robert. “Robots and Respect: Assessing the Case Against Autonomous Weapon Systems.” *Ethics & International Affairs* 30 no. 1 (2016): 93–116. doi:10.1017/S0892679415000647

Spelke, Elizabeth S. and Katherine D. Kinzler. Core Knowledge. *Developmental Science* 10, no. 1 (2007): 89-96.

Taggart, Kendall. “The Truth About The Trump Data Team That People Are Freaking Out About.” *BuzzFeed*. February 17, 2017. https://www.buzzfeed.com/kendalltaggart/the-truth-about-the-trump-data-team-that-people-are-freaking

Taleb, Nassim Nicholas. *The Black Swan: The Impact of the Highly Improbable*. London: Allen Lane, 2011.

Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Alfred A. Knopf, 2017.

Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. New York: Random House, 2015.

Torres, P. Agential Risks: A Comprehensive Introduction. *Journal of Evolution and Technology*, 26, 31-47 (2016). http://jetpress.org/v26.2/torres.pdf

Torres, P. *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Durham, North Carolina: Pitchstone Publishing, 2017.

USAMRIID. "Medical Management of Biological Casualties Handbook." 8th Edition. United States Army Medical Research Institute of Infectious Diseases, 2014. http://www.usamriid.army.mil/education/bluebookpdf/USAMRIID%20BlueBook%208th%20Edition%20-%20Sep%202014.pdf

Vinding, Magnus. *Reflections on Intelligence*. Copenhagen: Magnus Vinding, 2016.

Wallach, Wendell, and Colin Allen. "Framing Robot Arms Control." *Ethics and Information Technology* 15 no. 2 (2013): Kluwer Academic Publishers: 125–35. doi:10.1007/s10676-012-9303-0

Williamson, Jack. *With Folded Hands*. Reading: Fantasy Press, 1947.

Wilson, Grant. "Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law." *Virginia Environmental Law Journal* 31 no. 1 (2013): 307–64. http://lib.law.virginia.edu/lawjournals/sites/lawjournals/files/3.%20Wilson%20-%20Emerging%20Technologies.pdf

Yampolskiy, Roman V. "Leakproofing the Singularity: Artificial Intelligence Confinement Problem." *Journal of Consciousness Studies* 19 no. 1 (2012): 194–214. http://cecs.louisville.edu/ry/LeakproofingtheSingularity.pdf

Yampolskiy, Roman V. "Taxonomy of Pathways to Dangerous AI." *arXiv*, 2015. http://arxiv.org/abs/1511.03246

Yudkowsky, Eliezer. “Staring into the Singularity 1.2.5.” *Yudkowsky.net*, 1996. http://yudkowsky.net/obsolete/singularity.html

Yudkowsky, Eliezer. “Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.” *Singularity Institute for Artificial Intelligence* 15 (2001). https://intelligence.org/files/CFAI.pdf

Yudkowsky, Eliezer. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–45. New York: Oxford University Press, 2008. https://intelligence.org/files/AIPosNegFactor.pdf

Yudkowsky, Eliezer. “Hard Takeoff.” *LessWrong*, December 2, 2008. http://lesswrong.com/lw/wf/hard_takeoff/