How Does Artificial Intelligence Pose an Existential Risk?

By Karina Vold and Daniel R. Harris

1. Introduction

The idea that AI might one day threaten humanity has been around for some time. In 1863, the novelist Samuel Butler (1863 ,185) suggested that machines may one day hold "supremacy over the world and its inhabitants". By the mid-twentieth century, these concerns had left the realm of science fiction, as thinkers like Alan Turing (1951, 260) began to warn the public that we should expect intelligent machines to eventually "take control". Still, for many years, academics did not spill much ink over these concerns, even while Hollywood filmmakers ran with them, producing countless blockbusters based on this "AI takeover" scenario (think: *The Terminator* or *Battlestar Galactica*). Over the last decade or so, however, many leading academics and entrepreneurs have notably increased their attention to existential risks from AI. These concerns are, as we will see, more subtle than those depicted in crude Hollywood-produced AI takeover scenarios. Indeed, those depictions have largely misrepresented the concrete issues scholars are concerned with by overly focusing on anthropomorphic concerns of conscious AI systems deciding to destroy humans.

This renewed scholarly interest in AI safety has been spurred on in part by the recent deep learning revolution. This period is defined by major advances in the accomplishments of *deep neural networks*—artificial neural networks with multiple layers between the input and output layers—across a wide range of areas, including game-playing, speech and facial recognition, and image generation. Even with these breakthroughs though, the cognitive capabilities of current AI systems remain limited to domain-specific applications. Nevertheless, many researchers are alarmed by the speed of progress in AI and worry that future systems, if not managed correctly, could present an existential threat.

Despite the renewed interest in this concern, there remains substantial disagreement over both the nature and the likelihood of the existential threats posed by AI. Hence, our aim in this chapter is to explicate the main arguments that have been given for thinking that AI does pose an existential risk, and to point out where there are disagreements and weakness in these arguments. The chapter has the following structure: in §2, we will introduce the concept of existential risk, the sources of such risks, and how these risks are typically assessed. In §3–5, we will critically examine three commonly cited reasons for thinking that AI poses an existential threat to humanity: the control problem, global disruption from an AI "arms race", and the weaponization of AI. Our focus is on the first of these three, because it represents a kind of existential risk that is novel to AI as technology. While the latter two are equally important, they have commonalities with other kinds of technologies (e.g., nuclear weapons) discussed in the literature on existential risk, and so we will dedicate less time to them.

2. What Is an Existential Risk?

Many people believe that *existential risks* (henceforth, *Xrisks*) are the greatest threats facing humanity. And whilst there is much common ground amongst scholars about which scenarios constitute an Xrisk—the most commonly cited example is *extinction risks*[1]—there is not as much consensus on the precise definition of the concept (Beard et al., 2020; Torres, 2019). While most Xrisk scholars agree that a risk is existential if an adverse outcome would bring about human extinction, few endorse the narrower view that a risk is existential *only if* it would cause this outcome.[2] Most definitions of Xrisk are broader, including at times the risk of global civilizational collapse (Rees, 2003; Ó hÉigeartaigh, 2017); scenarios in which the technological and moral potential of humanity is "permanently and drastically" curtailed (Bostrom, 2002, 2013); and *suffering risks*, defined as cases in which "an adverse outcome would bring about severe suffering on an astronomical scale, vastly exceeding

[1] Extinction risks are those that directly cause the extinction of the human species or less directly lead to circumstances that cause our extinction (e.g., through habitat destruction). (See discussions in Matheny, 2007; Bostrom, 2013; and Cotton-Barratt et al., 2020.)

[2] Moynihan (2020) is the only example we found of someone using this narrow definition. C.f. Sotala & Gloor (2017) and Torres (2019), who both claim that the narrow definition is most common.

all suffering that has existed on Earth so far" (Sotala & Gloor, 2017, 389).

Xrisks are typically distinguished from the broader category of global catastrophic risks. Bostrom (2013), for example, uses two dimensions—scope and severity—to make this distinction. *Scope* refers to the number of people at risk, while *severity* refers to how badly the population in question would be affected (ibid, 16). Xrisks are at the most extreme end of both of these spectrums: they are *pan-generational* in scope (i.e., "affecting humanity over all, or almost all, future generations"), and they are the severest kinds of threats, causing either "death or a permanent and drastic reduction of quality of life" (ibid, 17). Perhaps the clearest example of an Xrisk is an asteroid impact on the scale of that which hit the Earth 66 million years ago, wiping out the dinosaurs (Schulte et al., 2010; Ó hÉigeartaigh, 2017). *Global catastrophic risks*, by way of contrast, could be either just as severe but narrower in scope, or just as broad but less severe. Some examples include the destruction of cultural heritage, thinning of the ozone layer, or even a large-scale pandemic outbreak (Bostrom, 2013). In this chapter, we will focus mostly on the least controversial category of Xrisks—extinction risks—but will also at times discuss some of the other scenarios mentioned.

2.1 Sources of Xrisk

For most of human history, the only source of Xrisks facing humanity were *natural causes*, such as an asteroid hitting Earth or a global pandemic (Bostrom, 2002). But the creation of the first atomic bomb in 1945 introduced a new source of existential threat to humanity, one that was *anthropogenic* in nature. But since then, humanity has created numerous other kinds of threats to our own existence, including human-caused climate change, global biodiversity loss, biological warfare, and threats from artificial intelligence, for example. In fact, it is widely thought that most Xrisks today are anthropogenic and that, as a result of these new threats, this current century is the riskiest one that humanity has ever faced (Rees, 2003; Bostrom, 2013; Ó hÉigeartaigh, 2017; Ord, 2020).

Not all of these threats pose straightforward Xrisks. Let's consider an extinction scenario to be the existential outcome in question, and then take nuclear fallout as an example. Today, the worldwide arsenal of nuclear weapons could lead to unprecedented death tolls and habitat destruction and, hence, it poses a clear global catastrophic risk. Still, experts assign a relatively low probability to human extinction from nuclear warfare (Martin, 1982; Sandberg & Bostrom, 2008; Shulman, 2012). This is in part because it seems more likely that extinction, if it follows at all, would occur *indirectly* from the effects of the war, rather than *directly*. This distinction has appeared in several discussions on Xrisks (e.g., Matheny, 2007, Liu et al., 2018; Zwetsloot & Dafoe, 2019), but it is made most explicitly in Cotton-Barratt et al. (2020, 6), who explain that a global catastrophe that causes human extinction can do so either *directly* by "killing everyone", or *indirectly*, by "removing our ability to continue flourishing over a longer period." A nuclear explosion itself is unlikely to kill *everyone* directly, but the resulting effects it has on the Earth could lead to lands becoming uninhabitable, in turn leading to a scarcity of essential resources, which could (over a number of years) lead to human extinction. Some of the simplest examples of *direct* risks of human extinction, by way of contrast, are "[i]f the entire planet is struck by a deadly gamma ray burst, or enough of a deadly toxin is dispersed through the atmosphere" (ibid, 6). What's critical here is that for an Xrisk to be *direct* it has to be able to reach *everyone*.

Much like nuclear fallout, the arguments for why and how AI poses an Xrisk are not straightforward. This is partly because AI is a general-purpose technology. It has a wide range of potential uses, for a wide range of actors, across a wide range of sectors. In this chapter, we are interested in the extent to which the use or misuse of AI can play a *sine qua non* role in Xrisk scenarios, across any of these domains. We are interested not only in current AI capabilities, but also in future (potential) capabilities. Depending on how the technology develops, AI could pose either a direct or indirect risk, although we make the case that direct Xrisks from AI are even more improbable than indirect ones. Another helpful way of thinking about AI risks is to divide them into accidental risks, structural risks, or misuse risks (Zwetsloot & Dafoe, 2019). In §3, we focus on *accidental risks*: threats arising from the system behaving in unintended ways. In §4, we turn to *structural risks*: threats arising from how the technology shapes the broader environment, especially in the political and military realms, in ways that can elevate risk. And finally, in §5, we examine potential *misuses* of AI.

Before moving on, it is worth noting that there are some significant methodological challenges that confront the study of Xrisks. Because events that constitute or precipitate an Xrisk are unprecedented, arguments to the effect that they pose such a threat must be theoretical in nature. Their rarity also makes it such that any speculations about how or when such events might occur are subjective and not empirically verifiable (Sagan, 1983; Matheny, 2007; Beard, et al. 2020), which makes such claims challenging to submit to standard forms of risk analysis (Ó hÉigeartaigh, 2017). Despite these challenges, however, it is still important to try to distinguish which extreme scenarios are actually plausible and worthy of further attention, even if they have an extremely low probability, as opposed to those that can be dismissed as science fiction (ibid, 3-4). Accordingly, our goal in this chapter is not to assign probabilities to arguments that AI poses an Xrisk, but rather to assess their theoretical nature.

3. The Control Problem Argument for Xrisk

The earliest line of thinking that AI poses an Xrisk warns that AI might become both powerful and indifferent to human values, leading to dangerous consequences for human beings. Despite it being a longstanding concern, the structure of this argument is rarely, if ever, explicitly laid out.[3] By presenting *the control problem argument for Xrisk* (henceforth CPAX) in this way, our aim is to capture what we understand to be the line of reasoning while also making the epistemic moves more explicit.

CPAX rests on two central theses: the Orthogonality Thesis and the Instrumental Convergence Thesis, both of which were first explicitly articulated by Bostrom (2012, 130-132; 2014).

> *Orthogonality Thesis:* The intelligent capacities of any system are logically independent from any goals the system might have.

[3] Bostrom (2014) and Russell (2019) each have a book-length defence of these issues, but neither lay out CPAX in an explicit way. The closest examples we found were Chalmers (2010) and Danaher (2015). Here, we draw from all of these sources.

> *Instrumental Convergence Thesis:* Almost any intelligent system is likely to converge upon certain instrumental (sub)goals.

We will discuss each of these theses, as well as the premises and central inferences of the argument (below) in §3.1–§3.4.

> P1. It is possible to build an AI system that has a decisive strategic advantage over all other forms of intelligence.
>
> P2. If an AI system has a decisive strategic advantage over human intelligence, then we may not be able to control that system.
>
> C1. It is possible to build an AI system that we are not able to control *(from P1 and P2)*.
>
> P3. The intelligent capacities of an AI system are logically independent from any goals the system might have (supported by the Orthogonality Thesis).
>
> C2. Therefore, it is possible to build an AI system that human beings are not able to control and that has goals that do not align with human values *(from C1 and P3)*.
>
> P4. AI systems are likely to converge upon certain instrumental (sub)goals that are inimical to human interests (supported by the Instrumental Convergence Thesis).
>
> C3. It is possible to build AI systems that pose an existential threat to humanity *(from C2 and P4)*.

This reconstruction of the argument is by no means uncontroversial, and we will discuss some of the disagreements and objections as we go through the argument.

3.1 Intelligence Explosion and Decisive Strategic Advantages

P1 states that it is possible to build an AI system that has a decisive strategic advantage over all other forms of intelligence (including human intelligence). Historically, CPAX was introduced as arising from an intelligence explosion that would lead to the creation of a superintelligent AI—a system that by definition has a decisive strategic advantage over human intelligence. More recently, some have argued that an intelligence explosion is not the only pathway to AI gaining a decisive strategic advantage. We will begin by explaining the pathway to a loss of control over AI (C1) from an intelligence explosion (in this section, §3.1) and consider some potential objections (§3.1.1). We will then, in §3.2, discuss P2 and some more contemporary takes on how C1 could result.

An *intelligence explosion* is a hypothetical event in which an AI system enters a rapid cycle of recursive self-improvement, whereby each new iteration creates a more intelligent version of itself, culminating in the creation of a superintelligence. Here, a *superintelligence* is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom 2014, 22). The concept of an intelligence explosion was first articulated by I.J. Good (1965, 33), who argued that an AI system whose intelligence exceeds humanity's in *all* intellectual activities would necessarily also exceed it in terms of designing machine intelligence. Hence, if such a system were initially engineered by humans, it would possess the capability to design a machine more intelligent than itself. The subsequent new iteration, being more intelligent than its predecessor, would by the same logic also be capable of designing a machine more intelligent than itself. If each new generation of AI were to utilize its improved design capability, an intelligence explosion would occur (Chalmers, 2010).

Importantly, an intelligence explosion need not begin with the creation of a machine with greater than human intelligence, as Good's argument suggests. In principle, it could be sparked via the creation of a more modest type of machine intelligence. Some might hold, for example, that an intelligence explosion merely requires a system with *artificial general intelligence*, where general intelligence is the ability to deploy the same core suite of cognitive resources to complete a wide range of different tasks (Shevlin et al., 2019). An even more modest possibility is that an intelligence explosion could spark from a mere *artificial narrow intelligence*, that is, a system that excels only at specific tasks and lacks the ability to use its resources to solve problems outside of its narrow domains.[4] Bostrom (2014, 29), for example, suggests that a system "capable of improving its own *architecture*", what he calls a "seed AI", would be a sufficient starting point. For example, DeepMind's AlphaZero, a current narrow AI system, has already shown the capacity to iteratively self-improve by repeatedly playing against itself. This illustrates how, under certain conditions, this process of recursive self-improvement might generate an intelligence explosion that begins from a mere narrow AI, in particular, any narrow AI system that enjoys a *decisive strategic advantage* (i.e., well above human level capacity) in some relevant domains, coupled with sufficient capacities for real-world modification.[5,6]

3.1.1 Objections to the Possibility of an Intelligence Explosion

In this section, we consider two objections to the possibility of an intelligence explosion.

Objection one: Why think that an AI system would recursively self-improve just because it had the capacity to do so?[7] Indeed, it seems logically possible that even if a system *could* design a more intelligent iteration of itself, that it would not take this action. More broadly, some have argued that superintelligent AI systems would be "inextricably unpredictable", and hence there is nothing that can be said regarding their potential goals (Cortese, 2014).

[4] This final possibility is more modest in the sense that it would require less advancement in current technology as a starting point for the argument. This is because, while we currently have many sophisticated narrow AI systems, which outperform humans in certain tasks (e.g. playing chess or go), we do not yet have any general AI systems, and many scholars hypothesize that we are a long ways from developing these (e.g. Dignum, 2019).

[5] Here we distinguish our use of the phrase "decisive strategic advantage" from that of Bostrom's (2014, 78), who defines it as "a level of technological and other advantages sufficient to enable it to achieve complete world domination".

[6] AlphaZero, to be sure, could not spark an intelligence explosion, as it has neither a decisive strategic advantage *in a relevant domain* nor does it have sufficient capacities for real-world modification.

[7] We have bracketed those cases in which an AI system is deliberately engineered to recursively self-improve, as it straightforwardly follows that it would do so under these conditions.

Reply: In order to make meaningful predictions regarding what a sufficiently advanced AI would do (e.g., whether it would recursively self-improve), we need a method of identifying what goals it might have. Yet, the Orthogonality Thesis (which we discuss in more detail below) holds that the intelligent capacities of an AI system are logically independent from any goals the system might have. If true, the range of possible goals that an AI system could have is enormous. In light of this, how do we make predictions? Omohundro (2008, 1) argues that any sufficiently advanced AI is likely to have several basic "drives", or "tendencies which will be present unless explicitly counteracted". Bostrom (2014, 109) similarly argues that we can make robust inferences regarding the sub-goals of almost any intelligent agent by appealing to the thesis of *instrumental convergence*:

> "Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents."

Both Omohundro and Bostrom identify self-improvement as one of the "basic drives" or "instrumental values" (respectively) that a system would pursue. They also share the same general reasoning: improvements in rationality and intelligence are likely to be pursued by a wide variety of intelligent agents insofar as they tend to improve an agent's decision-making capabilities, thereby increasing the likelihood of that agent realizing whatever final objective it has (Bostrom 2014, 111). Taken together, the Bostrom-Omohundro thesis suggests that an AI system would have instrumental reasons to undergo a process of recursive self-improvement, especially cognitive self-improvement, and thus would be driven to do so.

Objection two: Despite the aforementioned reasons for thinking that a sufficiently intelligent AI system would be motivated to recursively self-improve, there are some who question whether an intelligence explosion is an inevitable outcome of the creation of such a system. Yoshua Bengio, for example, speculates that for mathematical and computational reasons there may be a "*wall-of-complexity*" that confronts all forms of intelligence "due to exponentially growing complexities" and that limits the capacities of any intelligent agent (quoted in Sofge, 2015). He speculates that this "wall" might (partly) explain why animals with bigger brains than ours are not more intelligent than us. Meanwhile, Chalmers (2010, 19-22) argues that a number of obstacles could arise that forestall an intelligence explosion. Among these are what he terms "*manifestation obstacles*"—difficulties which obstruct self-amplifying capabilities from developing.[8] He further subdivides these into two types of defeaters: *motivational* and *situational*. Motivational defeaters include disinclination and active prevention. For example, an AI system might discover that self-improvement(s) come at a cost which outweighs the associated gains, as would be the case if it turns out that improvements in intelligence have diminishing returns.[9] Alternatively, we might design the AI system to lack the motivation to self-improve, or to have a contrary motivation which supersedes it (ibid).[10] As for situational defeaters, these include unfavourable circumstances, such as a limitation to the availability of resources necessary for cognitive upgrades (ibid).

Reply: A defender of CPAX has to make the case that it is unlikely for these defeaters to be present, or at the very least, that there is a non-zero chance they will be absent. But, we think Chalmers (2010, 19) is right in writing that there are no "knockdown arguments against any of these obstacles". As with the previous objection, both sides of this issue confront the same challenge: that it is difficult, perhaps impossible, to predict what the motivations of a future advanced AI system could be.[11]

[8] Chalmers also identifies two other types of obstacles—*structural obstacles* and *correlation obstacles*.

[9] See further discussion in Russell (2019).

[10] By "motivations", we don't mean to attribute intentional agency, but rather "agency" in a kind of Dennettian sense (i.e., following the intentional stance). We mean motivations as something like "sub-goals", or the strategies a system will undertake to achieve its final goals.

[11] Consider another scenario, discussed by Tegmark (2014) and Häggström (2019), in which an advanced AI is programmed to have a meaningless or "undefined" goal. What would happen in this case? Häggström predicts that all instrumental goals would also

Ultimately, an intelligence explosion is certainly not an inevitable outcome, but it also is not an impossible one. Chalmers (2010, 22) takes a similar position, saying that it is "far from obvious that there will be defeaters", and in their absence, the outcome of an intelligence explosion would be the creation of a superintelligence.

3.2 Losing Control (P2 and C1)

If we take as true the supposition that humanity's control over other Earth-bound species is largely the result of our comparative intelligence advantage, then the emergence of an AI system with a decisive strategic advantage over human intelligence should give cause for concern. Consider the current power structure between human beings and gorillas. Bostrom (2014, vii) argues that due to our advantage in general intelligence "the fate of the gorilla now depends more on us humans than on the gorillas themselves". Russell (2019, 134) makes a similar point, noting that by virtue of our intelligence, gorillas "essentially have no future beyond that which we deign to allow". Both authors, among others (e.g., Hawking, 2018), raise the worry that, in much the same way, the destiny of humanity could be dictated by the actions of a superintelligent AI. This "*gorilla problem*", as Russell terms it, is the idea behind P2, which states that if an AI system has a decisive strategic advantage over human intelligence, then we may not be able to control that system.

There are a few reasons for thinking that we will not be able to control systems that are more intelligent than us. Among these are the opacity and unpredictability of such systems. A good example of these features in a non-critical domain is DeepMind's AlphaGo—which in 2016 beat the 18-time world champion of Go, Lee Sedol. In the games against Sedol, the system sometimes made moves that proved advantageous but that both the engineers of the system and human Go experts alike did not foresee and struggled to interpret (Silver et al., 2016; Kohs, 2017). The opaqueness and unpredictability of these systems makes them potentially dangerous—if even experts cannot predict or interpret how a system will behave, it could run amok sooner than we realize and before we have a chance to intervene. Hence, it follows from the first two premises that we could risk losing control over what a

then become pointless, in which case "all predictions" from the instrumental convergence thesis "collapse".

strategically advantaged AI system does (C1). The possibility of this occurring is often referred to as *the control problem*. But, for this problem to pose an Xrisk, a few more premises are needed.

Before continuing on, however, we should note two things. First, notice that nothing so far, or going forward, in the discussion of CPAX relies on the idea that an AI system would need human-like motivations. This is important to clarify because some critics object that CPAX relies on erroneous anthropomorphic assumptions about AI (e.g., Andrew Ng (Williams, 2015) and Yann LeCun (Wakefield, 2015)). But the central idea so far is that an AI might cause harm for instrumental reasons, or that it might have little regard for human life or have that regard outweighed by other concerns. Hawking (2018, 188) explains:

> "[T]he real risk with AI isn't malice, but competence. … You're probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green-energy project and there's an anthill in the region to be flooded, too bad for the ants. Let's not place humanity in the position of those ants."

In other words, CPAX need not assume that an AI would develop malicious aims or choose to destroy humanity because of human-like emotions, such as disgust, revenge, or anger. It also does not assume the AI in question would be conscious, or even that it would become (non-consciously) motivated to harm or exterminate human beings.

The second thing to note is that, while early concerns around the control problem (that is, premises P1 and P2, leading up to C1) focused on the possibility of an intelligence explosion, more recent discussions have moved away from this scenario. In other words, it has been argued that a loss of control (i.e., C1) could result without an intelligence explosion and without the emergence of a superintelligent AI. An AI system might not require an internal "drive" to self-improve, for example, if human beings are incentivized to aid its improvement (Drexler, 2019). It also may not need to reach the level of superintelligence in order to pose an Xrisk (ibid). Russell (2019, 137) explains that humanity has thus far been protected from the "potentially catastrophic consequences" of AI because of the limited intelligent capacities of current AI systems and their limited abilities to bring about changes in the real world (most systems operate in virtual worlds or lab environments). But as narrow AI systems become more cognitively sophisticated and are given more capacity to directly affect or modify the world, they could pose an Xrisk as long as their narrow domain is critical enough (e.g., controlling stock markets or military decision-making) and their interests are inimical to those of humans.

3.3 The Orthogonality Thesis (P3 and C2)

In one of the earliest explanations of the control problem, Norbert Weiner (1960, 1358) worried there could be dangerous outcomes if a powerful AI system that we lacked control over were to operate with an incorrect objective: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere… we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colourful imitation of it." Consider an example that Russell (2019, 138) gives of a machine tasked with solving environmental problems:

> "[Y]ou might ask the machine to counter the rapid acidification of the oceans that results from higher carbon dioxide levels. The machine develops a new catalyst that facilitates an incredibly rapid chemical reaction between ocean and atmosphere and restores the oceans' pH levels. Unfortunately, a quarter of the oxygen in the atmosphere is used up in the process leaving us to asphyxiate slowly and painfully. Oops."

In this case, the Xrisk arises from incidental safety issues that fall out of misaligned objectives. Bostrom (2014, 97) offers a similar example of an AI that might "tile all of the Earth's surface with solar panels, nuclear reactors, supercomputing facilities with protruding cooling towers, space rocket launchers, or other installations whereby the AI intends to maximize the long-term cumulative realization of its values." Once again, the threat to humanity here is essentially a side effect of the widespread habitat destruction that would ensue.[12] Because AI is a general-purpose technology, the misalignment problem could arise in

[12] These are two of many possible hypothetical scenarios of how powerful AIs with misaligned objectives could be Xrisks. See Shanahan (2015) and Tegmark (2017) for further examples.

many different domains, though Xrisks seem most likely to occur in domains with global impact (because they are by definition pan-generational in scope).

For some, the idea that a superintelligent machine would pursue such narrow goals without regard to broader consequences seems improbable, if not "self-refuting" (Pinker, 2019; other critics include Loosemore, 2012; Chorost, 2016; Metzinger, 2017). Loosemore (2012) has dubbed it the "fallacy of dumb superintelligence". While Chorost (2016), responding to Bostrom's scenario, argues that "By the time [a superintelligent AI] is in a position to imagine tiling the Earth with solar panels, it'll know that it would be morally wrong to do so." Metzinger (2017) makes this same claim, arguing that because a superintelligence would be better than human beings at moral cognition it would also be benevolent.[13] But the possibility that a highly intelligent artificial agent could act in ways that are malevolent or misaligned with the values of its designers is meant to follow from Bostrom's *Orthogonality Thesis*, which maintains that the intelligent capacities of any system are logically independent from any goals the system might have. By "intelligent capacities" here, Bostrom (2012, 74) means the capacities related to *instrumental rationality*, e.g., "skill at prediction, planning, and means-ends reasoning in general". This Orthogonality Thesis is meant to apply quite broadly to any intelligent system, including humans. And Hume's (1739) longstanding is-ought problem lends support for the idea: if one cannot infer normative statements from descriptive ones, then however intelligent a system is, it may never arrive at any moral facts (Bostrom, 2012, 74; Armstrong, 2013). The Orthogonality Thesis supports the third premise of CPAX, and with C1, it leads by conjunction to the second conclusion, C2: that it is possible to build an uncontrollable AI system that has goals that do not align with human values.

3.3.1 The Value Alignment Problem

Arguably, one way to avoid C2 would be to program the AI to ensure that it is benevolent or that its values are reliably aligned with our own. After all, humans have the advantage of being the ones who build the system and determine its initial goals (Bostrom, 2014; Russell, 2019).

[13] We could not find any scholarly publications in which Loosemore or Metzinger make these points, so in both cases we cite online articles.

Unfortunately, it's not that simple. This is widely known as the *value alignment problem* (see Gabriel & Ghazavi, this volume). The problem is hard for many reasons.

First, there is the *issue of identifying human values*. Human beings are often confused and conflicted about our own values, and different cultures seem to have wide variation between their respective values.[14] While pervasive moral disagreement does not necessarily imply value relativism, it does complicate the challenge of trying to build machines that align with "our" values, as it opens the door to value pluralism (the view that there are many different and sometimes irreducible moral values). While a value monist has the challenge of identifying the one value (e.g., happiness or pleasure) that all other values reduce to, the value pluralist has the problem of identifying and implementing the complete set of irreducible values. This is a real challenge for programming; as Russell (2019, 139) notes, one of the most common forms of value misalignment comes from an incomplete articulation of values, that is, from omitting something human beings care about from the objective imbued into the system. Furthermore, value pluralists tend to believe that there are at least some, and perhaps many, unresolvable moral dilemmas that result from a conflict between incommensurable values (further discussions in Cave et al., 2019; Baum, 2020).

A second problem is that even if we can identify some acceptable set of human values or objectives, we are rather prone to misstating these. This is sometimes known as the "*King Midas problem*" (Russell, 2019).[15] In wishing that everything he touched should turn to gold, King Midas thought he knew what he wanted, but he didn't really want his wife or his breakfast to turn to gold. The folklore illuminates the issue of "*value fragility*"—if an AI system gets our values even slightly wrong, it could lead to disastrous outcomes. Hence, the more we rely on powerful autonomous systems, the more important it will be for us to specify their

[14] We say "seem" here because of the work in cross-cultural psychology over the last two decades which tries to show that human societies that seem to vary widely in their values in fact share some basic set of normative commitments (or values), even though the shared set may be interpreted or applied differently (e.g. Borg et al., 2019 and Christians, 2019).

[15] Also sometimes referred to as the "specification problem", the "genie problem", or the "Sorcerer's Apprentice problem".

goals with great care, ensuring that we express our objectives correctly and completely. Yet, most goals that are easy to specify will not capture the context-specific complexities of human objectives in the real world.[16] And indeed, AI systems frequently find ways to maximize their reward functions with unintended behaviours—what Bostrom (2014, 120-124) calls "*perverse instantiations*". A nice example is given by Russell & Norvig (2010, 37), who imagine a vacuum robot whose performance is measured by the amount of dirt it cleans up. The optimal learned policy causes the robot to repeatedly dump and clean up the same dirt, which is obviously not what the designer of a vacuum intends her machine to do. It is not obvious, however, whether these examples of "specification gaming" in current systems should count as evidence that future systems with more advanced intelligence are also likely to behave in these ways.[17]

A third reason that the value alignment problem is challenging is that our own values—assuming we could identify and perfectly articulate them—are not perfect. Human beings are far from reliably human-friendly. If superintelligent machines merely aim to achieve our own standards of "human friendliness" or "friendliness to other life forms", we may not be very well off (Price & Vold, 2018). Indeed, we may find ourselves living amongst superintelligent systems that amplify our own fallible, inconsistent, and complacent moral natures. Let's call this the *problem of human moral imperfection*. A related problem emerges from the need to accommodate moral progress. Even if we can find a way to build machines that align with (only) the better parts of our current values, we would not want AI systems to codify these values in a way that prevents moral progress. After all, one does not have to look far into human history to see how much our values have progressed (ibid).

3.4 Concluding CPAX (P4 and C3)

P4 of CPAX states that AI systems are likely to converge upon certain instrumental goals that are inimical to human interests, a premise supported by the Instrumental Convergence Thesis (above). The idea

[16] See Cantwell Smith (2019) for an argument that this capacity for "deep contextual awareness" in our ethical judgements (i.e. to discern what norms or values apply within a specific fine-grained context)—a capacity that is central to virtue ethics—is computationally intractable.

[17] Thanks to Matthijs Maas for raising this point in discussion.

behind P4 is that some of the goals that intelligent systems are likely to converge upon will put those systems (e.g., an AI and humans) at odds with each other. As just one example, both Omohundro and Bostrom identify *resource acquisition* as a basic drive and an instrumental convergent value, respectively. In Omohundro's view, "[a]ll computation and physical action requires the physical resources of space, time, matter, and free energy", and hence, "almost any goal can be better accomplished by having more of these resources" (2008, 491). Bostrom (2014, 114-116) argues that, for this reason, it is likely that "an extremely wide range of possible final goals" would generate "the instrumental goal of unlimited resource acquisition".[18] Perhaps the most widely discussed example of this problem is Bostrom's (2014, 123) paperclip maximizer—a superintelligent AI system that has the goal of maximizing the production of paperclips. The system finds any means necessary of producing more paperclips, including securing any resources necessary for that purpose. With sufficient capacities to modify the world, soon enough the system could co-opt much of the Earth's natural resources, including those needed for the survival of humanity, all for the purposes of paperclip production. The example is meant to show that even with good intentions and fairly innocuous goals, we could end up with AI systems inadvertently acting in ways that are inimical to human values. Another related concern is that human beings not only require resources to survive and flourish—we are also resources ourselves. In the *Matrix Trilogy*, for example, the AI system turns humans into an energy source to power itself.

Both of these examples demonstrate how C3, which states that it is possible to build AI systems that pose an Xrisk, is meant to follow from C2 and P4. But, here we again face both a dearth of critical discussion and of compelling examples. Indeed, because most of the CPAX scenarios offered by leading defenders (e.g., Bostrom, 2014; Russell, 2019) are so "bemusing",[19] one is inclined to simply dismiss them as belonging to science fiction. More charitably, one could see them as toy examples, meant to illustrate a broader concern about misaligned powerful AI systems—a concern that we've tried to outline in more detail. While any specific scenario might be dismissed as unrealistic, the

[18] Similar arguments for instrumental convergence around resource acquisition can be found in Tegmark (2017, 266) and Russell (2019).

[19] As Leslie (2019) describes them in a critical review.

broader concern remains *possible*. It is a (very) low probability risk that hinges critically on certain assumptions about (a) the possible motivations of an advanced AI system and (b) the potential capacities that such a system could possess to bring about critical changes in the world (i.e., capacities for direct world modification). Notice as well that the argument only supports the (low probability) possibility of AI posing an *indirect Xrisk*. None of the scenarios discussed suggest that an AI system would, for example, *directly* eliminate the whole of the population.

In closing, it's worth noting that many think that the best way to mitigate the risks of the control problem is to ensure that we build the initial conditions of the system in a way that aligns with human values, thereby avoiding P4 of CPAX. For further reading, we direct readers to the subfields of AI safety engineering and machine ethics, both of which have taken on the goal of trying to find technical solutions for building ethically aligned systems (Cave et al., 2019).

4. AI Race Dynamics, Global Disruption & Xrisk

While the earliest arguments for AI Xrisk focused on control problem scenarios that were based around hypothetical advanced AI systems, a set of recent arguments centre on a more immediate, and practically grounded, set of issues. One of these is the growing concern that advanced AI could confer significant strategic advantages to its possessors, and correspondingly, whether an AI race could emerge between powerful actors in pursuit of this technology (Dafoe, 2018; Cave & Ó hÉigeartaigh, 2018; Bostrom, 2014).[20] In §4.1 and §4.2 we discuss two associated sources of *structural risks* that could pose *indirect Xrisks* should such an AI race dynamic arise: the first is that it could disincentivize researchers from investing in AI safety, and the second is that it could spark military conflict between AI competitors. A related issue gaining attention is the impact AI could have on global strategic stability. We take this up in §4.3, focusing on its capacity to destabilize nuclear deterrence, and thereby potentially contributing to military

[20] Geist (2016) takes a stronger view, arguing that the world superpowers are already locked in a particular type of racing dynamic—an AI arms race.

conflict escalation.

4.1 AI Race Dynamics: Corner-cutting Safety

An AI race between powerful actors could have an adverse effect on AI safety, a subfield aimed at finding technical solutions to building "advanced AI systems that are safe and beneficial" (Dafoe, 2018, 25; Cave & Ó hÉigeartaigh, 2018; Bostrom, 2017; Armstrong et al., 2016; Bostrom, 2014). Dafoe (2018, 43), for example, argues that it is plausible that such a race would provide strong incentives for researchers to trade-off safety in order to increase the chances of gaining a relative advantage over a competitor.[21] In Bostrom's (2017) view, competitive races would disincentivize two options for a frontrunner: (a) slowing down or pausing the development of an AI system and (b) implementing safety-related performance handicapping. Both, he argues, have worrying consequences for AI safety.

(a) Bostrom (2017, 5) considers a case in which a solution to the control problem (C1) is dependent upon the components of an AI system to which it will be applied, such that it is only possible to invent or install a necessary control mechanism after the system has been developed to a significantly high degree. He contends that, in situations like these, it is vital that a team is able to pause further development until the required safety work can be performed (ibid). Yet, if implementing these controls requires a substantial amount of additional time and resources, then in a tight competitive race dynamic, any team that decides to initiate this safety work would likely surrender its lead to a competitor who forgoes doing so (ibid). If competitors don't reach an agreement on safety standards, then it is possible that a "*risk-race to the bottom*" could arise, driving each team to take increasing risks by investing minimally in safety (Bostrom, 2014, 247).

(b) Bostrom (2017, 5-6) also considers possible scenarios in which the "mechanisms needed to make an AI safe reduces the AI's effectiveness". These include cases in which a safe AI would run at a considerably slower speed than an unsafe one, or those in which implementing a safety mechanism necessitates the curtailing of an AI's capabilities (ibid). If the AI race were to confer large strategic and

[21] Armstrong's et al. (2016) model of an AI race dynamic supports this claim.

economic benefits to frontrunners, then teams would be disincentivized from implementing these sorts of safety mechanisms. The same, however, does not necessarily hold true of less competitive race dynamics; that is, ones in which a competitor has a significant lead over others (ibid). Under these conditions, it is conceivable that there could be enough of a time advantage that frontrunners could unilaterally apply performance handicapping safety measures without relinquishing their lead (ibid).

It is relatively uncontroversial to suggest that reducing investment in AI safety could lead to a host of associated dangers. Improper safety precautions could produce all kinds of unintended harms from misstated objectives or from specification gaming, for example. They could also lead to a higher prevalence of AI system vulnerabilities which are intentionally exploited by malicious actors for destructive ends, as in the case of adversarial examples (see Brundage et al., 2018). But does AI safety corner-cutting reach the threshold of an *Xrisk*? Certainly not directly, but there are at least some circumstances under which it would do so indirectly. Recall that Chalmers (2010) argues there could be defeaters that obstruct the self-amplifying capabilities of an advanced AI, which could in turn forestall the occurrence of an intelligence explosion. Scenario (a) above made the case that a competitive AI race would disincentivize researchers from investing in developing safety precautions aimed at preventing an intelligence explosion (e.g., motivational defeaters). Thus, in cases in which an AI race is centred on the development of artificial general intelligence, a seed AI with the capacity to self-improve, or even an advanced narrow AI (as per §3.1), a competitive race dynamic could pose an *indirect Xrisk* insofar as it contributes to a set of conditions that elevate the risk of a control problem occurring (Bostrom, 2014, 246; 2017, 5).

4.2 AI Race Dynamics: Conflict Between AI Competitors

The mere narrative of an AI race could also, under certain conditions, increase the risk of military conflict between competing groups. Cave & Ó hÉigeartaigh (2018) argue that AI race narratives which frame the future trajectory of AI development in terms of technological advantage could "increase the risk of competition in AI causing real conflict (overt or covert)". The militarized language typical of race dynamics may encourage competitors to view each other "as threats or even enemies" (ibid, 3).[22] If a government believes that an adversary is pursuing a strategic advantage in AI that could result in their technological dominance, then this alone could provide a motivating reason to use aggression against the adversary (ibid; Bostrom, 2014). An AI race narrative could thus lead to crisis escalation between states. However, the resulting conflict, should it arise, need not directly involve AI systems. And it's an open question whether said conflict would meet the Xrisk threshold. Under conditions where it does (perhaps nuclear war), the contributions of AI as a technology would at best be *indirect*.

4.3 Global Disruption: Destabilization of Nuclear Deterrents

Another type of crisis escalation associated with AI is the potential destabilizing impact the technology could have on global strategic stability;[23] in particular, its capacity to destabilize nuclear deterrence strategies (Giest & Lohn, 2018; Rickli, 2019; Sauer, 2019; Groll, 2018; Zwetsloot & Dafoe, 2019). In general, deterrence relies both on states possessing secure second-strike capabilities (Zwetsloot & Dafoe, 2019) and, at the same time, on a state's inability to locate, with certainty, an adversary's nuclear second-strike forces (Rickli, 2019). This could change, however, with advances in AI (ibid). For example, AI-enabled surveillance and reconnaissance systems, unmanned underwater vehicles, and data analysis could allow a state to both closely track and destroy an adversary's previously hidden nuclear-powered ballistic missile submarines (Zwetsloot & Dafoe, 2019). If their second-strike nuclear capabilities were to become vulnerable to a first strike, then a pre-emptive nuclear strike would, in theory, become a viable strategy under certain scenarios (Giest & Lohn, 2018).

In Zwetsloot & Dafoe's (2019) view, "the fear that nuclear systems could be insecure would, in turn, create pressures for states—including defensively motivated ones—to pre-emptively escalate during a crisis". What is perhaps most alarming is that the aforementioned AI systems need not actually exist to have a destabilizing impact on nuclear deterrence (Rickli, 2019; Groll, 2018; Giest & Lohn, 2018). As Rickli

[22] Here Cave & Ó hÉigeartaigh discuss Huysmans (2006).

[23] Giest & Lohn (2018, 10) define strategic stability as existing "when adversaries lack a significant incentive to engage in provocative behavior".

(2019, 95) points out, "[b]y its very nature, nuclear deterrence is highly psychological and relies on the perception of the adversary's capabilities and intentions". Thus, the "simple misperception of the adversary's AI capabilities is destabilizing in itself" (ibid). This potential for AI to destabilize nuclear deterrence represents yet another kind of indirect global catastrophic, and perhaps even existential, risk insofar as the destabilization could contribute to nuclear conflict escalation.

5. Weaponization of AI

Much like the more recent set of growing concerns around an AI arms race, there have also been growing concerns around the weaponization of AI. We use "weaponization" to encompass many possible scenarios, from malicious actors or a malicious AI itself, to the use of fully autonomous lethal weapons. And we will discuss each of these possibilities in turn. In §5.1 we discuss malicious actors and in §5.2 we discuss lethal autonomous weapons. We have combined this diverse range of scenarios for two reasons. First, while the previous Xrisk scenarios discussed (CPAX and an AI race) could emerge without malicious intentions from anyone involved (e.g., engineers or governments), the scenarios we discuss here do for the most part assume some kind of *malicious intent* on the part of some actor. They are what Zwetsloot & Dafoe (2019,) call a *misuse* risk. Second, the threats we discuss here are not particularly unique to AI, unlike those in previous sections. The control problem, for example, is distinctive of AI as a technology, in the sense that the problem did not exist before we began building intelligent systems. On the other hand, many technologies can be weaponized. In this respect, AI is no different. It is because AI is potentially so powerful that its misuse in a complex and high impact environment, such as warfare, could pose an Xrisk.

5.1 Malicious Actors

In discussing CPAX, we focused on *accidental risk* scenarios—where no one involved wants to bring about harm, but the mere act of building an advanced AI system creates an Xrisk. But AI could also be deliberately *misused*. These can include things like exploiting software vulnerabilities, for example, through automated hacking or adversarial examples; generating political discord or misinformation with synthetic media; or initiating physical attacks using drones or automated weapons (see Brundage et al., 2018). For these scenarios to reach the threshold of Xrisk (in terms of 'scope'), however, a beyond catastrophic amount of damage would have to be done. Perhaps one instructs an AI system to suck up all the oxygen in the air, to launch all the nuclear weapons in a nation's arsenal, or to invent a deadly airborne biological virus. Or perhaps a lone actor is able to use AI to hack critical infrastructures, including some that manage large-scale projects, such as the satellites that orbit Earth. It does not take much creativity to drum up a scenario in which an AI system, if put in the wrong hands, could pose an Xrisk. But the Xrisk posed by AI in these cases is likely to be *indirect*—where AI is just one link in the causal chain, perhaps even a distal one. This involvement of malicious actors is one of the more common concerns around the weaponization of AI. Automated systems that have war-fighting capacities or that are in anyway linked to nuclear missile systems could become likely targets of malicious actors aiming to cause widespread harm. This threat is serious, but the theoretical nature of the threat is straightforward relative to those posed in CPAX, for example.

One further novel outcome of AI would be if the system itself malfunctions. Any technology can malfunction, and in the case of an AI system that had control over real-world weapons systems the consequences of a malfunction could be severe (see Robillard, this volume). We'll discuss this potential scenario a bit more in the next section. A final related possibility here would be for the AI to itself turn malicious. This would be unlike any other technology in the past. But since AI is a kind of intelligent agent, there is this possibility. Cotton-Barratt et al. (2020), for example, describe a hypothetical scenario in which an intelligence explosion produces a powerful AI that wipes out human beings in order to pre-empt any interference with its own objectives. They describe this as a *direct* Xrisk (by contrast, we described CPAX scenarios as *indirect*), presumably because they describe the AI as *deliberately* wiping out humanity. However, if the system has agency in a meaningful sense, such that it is making these kinds of deliberate malicious decisions, then this seems to assume it has something akin to consciousness or strong intentionality. In general we are far from developing anything like artificial consciousness and this is not to say that these scenarios should be dismissed altogether, but many experts agree that there are serious challenges confronting the possibility of AI possessing these cognitive capacities (e.g., Searle, 1980; Koch and Tonini, 2017; Koch, 2019; Dehaene et al., 2017).

5.2 Lethal Autonomous Weapons

One other form of weaponization of AI that is sometimes discussed as a potential source of Xrisk are lethal autonomous weapons systems (LAWS). LAWS include systems that can locate, select, and engage targets without any human intervention (Roff, 2014; Russell, 2015; Robillard, this volume). Much of the debate around the ethics of LAWS has focused on whether their use would violate human dignity (Lim, 2019; Rosert & Sauer, 2019; Sharkey, 2019), whether they could leave critical responsibility gaps in warfare (Sparrow, 2007; Robillard, this volume), or whether they could undermine the principles of just war theory, such as noncombatant immunity (Roff, 2014), for example. These concerns, among others, have led many to call for a ban on their use (FLI ,2017). These concerns are certainly very serious and more near term (as some LAWS already exist) than the speculative scenarios discussed in CPAX. But do LAWS really present an *Xrisk*? It seems that if they do, they do so *indirectly*. Consider two possible scenarios.

(a) One concern around LAWS is that they will ease the cost of engaging in war, making it more likely that tensions between rival states rise to military engagement. In this case, LAWS would be used as an instrument to carry out the ends of some malicious actor. This is because, for now, humans continue to play a significant role in directing the behaviour of LAWS, though it is likely that we will see a steady increase in the autonomy of future systems (Brundage et al., 2018). Now, it could be that this kind of warfare leads to Xrisks, but this would require a causal chain that includes political disruption, perhaps failing states, and widespread mass murder. None of these scenarios are impossible, of course, and they present serious risks. But we have tried to focus this chapter on Xrisks that are *novel* to AI as a technology and, even though we view the risks of LAWS as extremely important, they ultimately present similar kinds of risks as nuclear weapons do. To the extent that LAWS have a destabilizing impact on norms and practices in warfare, for example, we think that scenarios similar to those discussed in §4.3 are possible—LAWS might escalate an ongoing crisis, or moreover, the mere perception that an adversary has LAWS might escalate a crisis.

(b) A second scenario, described by Geoffrey Hinton, is that killer drones, equipped with explosives and deep learning neural net technology, could (somehow) learn to function independently of their human controllers (Robinson, 2016), and the system could then go on a rampage and destroy humanity. The bracketed "somehow" here is a critical piece of the story. Perhaps the control system has been hacked, in which case we are back to the malicious actor scenario described in §5.1. Or perhaps there is a malfunction, of the sort also described in §5.1. In this latter case, the malfunction could manifest in the form of a "hard takeoff" in which the system undergoes rapid recursive self-improvement (unintended by the designers) and then develops goals that are inimical to human interests. In such a case, we would be at the start of an intelligence explosion and would confront the kind of Xrisk already characterized by CPAX (§3). Our only point here is that upon closer examination, it's hard to see how this scenario looks distinct from ones previously discussed. Hence, the weaponization of AI can pose an indirect Xrisk in several different ways. In general, the more control an automated system has over weaponized systems that can cause real-world destruction, the greater risk there is of that system becoming a target for attack by malicious actors or of there being greater harm due to any accidental system malfunction.

6. Conclusion

Humanity is facing an increasing number of existential threats, many of which are of our own creation. Thankfully, there are also an increasing number of scholars, from a wide range of fields, studying the nature of these risks and strategizing how to mitigate them. But the field of Xrisk studies is still relatively young. There are significant debates being had over how to define the concept of Xrisk, how to understand its sources, and what methodologies should be used to assess these risks. When it comes to Xrisks from AI, these debates continue. Early concerns around AI Xrisks focused on the possibility of an intelligence explosion and the subsequent pathway to a scenario in which a powerful superintelligent AI has misaligned objectives from humanity. These concerns have not gone away, but they have evolved over time. This chapter has provided an up-to-date critical survey of these arguments, both old and new, looking at different foreseeable pathways towards AI Xrisk, possible global disruptions resulting from the emergence of an AI race dynamic between nations, and the weaponization of AI. In particular, we have tried to make the structures of each of these concerns more explicit, such that readers can begin to critically engage with them.

Acknowledgements

We thank our colleagues from the University of Cambridge's Centre for the Study of Existential Risk who contributed to these ideas, including helpful input from Haydn Belfield, Martina Kunz, Matthijs Maas, Sean Ó hÉigeartaigh, Jaime Sevilla, Catherine Stinson, and Jess Whittlestone. We also thank our two anonymous reviewers for their invaluable comments. KV was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067. DH was supported by a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada.

Works cited

Armstrong, S., 2013. 'General purpose intelligence: Arguing the orthogonality thesis'. Analysis and Metaphysics 12: 68–84.

Armstrong, S., Bostrom, N. and C. Shulman. 2016. 'Racing to the Precipice: A Model of Artificial Intelligence Development'. AI & Society 31: 201–206.

Baum, S. D., 2020. 'Social choice ethics in artificial intelligence'. AI & Society 35: 165–176.

Beard, S., Rowe, T., and J. Fox. 2020. 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards'. Futures 115: 102469.

Borg, I., Hermann, D., Bilsky, W. *et al.* (2019). Do the PVQ and the IRVS scales for personal values support Schwartz's value circle model or Klages' value dimensions model?. *Meas Instrum Soc Sci* **1,** 3 (2019). https://doi.org/10.1186/s42409-018-0004-2

Bostrom, N., 2002. 'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards'. Journal of Evolution and Technology 9 (1).

Bostrom, N., 2012. 'The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents'. Minds and Machines 22 (2): 71–78.

Bostrom, N., 2013. 'Existential Risk Prevention as Global Priority'. Global Policy 4 (1): 15–31.

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

Bostrom, N., 2017. 'Strategic Implications of Openness in AI Development'. Global Policy 8 (2): 135–148.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., et al. 2018. 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation'. [https://maliciousaireport.com]

Butler, S., 1863. 'Darwin among the Machines'. Christchurch in the Press Newspaper, 13 June.

Cantwell Smith, B. 2019. The Promise of Artificial Intelligence: Reckoning and Judgment. Cambridge: MIT Press.

Cave, S., Nyrup, R., Vold, K., and A. Weller. 2019. 'Motivations and Risks of Machine Ethics'. Proceedings of the IEEE 107 (3): 562–574.

Cave, S. and S. Ó hÉigeartaigh. 2018. 'An AI race for strategic advantage: rhetoric and risks'. Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, New Orleans: 36–40.

Chalmers, D. J., 2010. 'The Singularity: A Philosophical Analysis'. Journal of Consciousness Studies 17 (9-10): 7–65.

Chorost, M., 2016. 'Let Artificial Intelligence Evolve'. Slate, 18 April. [https://slate.com/technology/2016/04/the-philosophical-argument-against-artificial-intelligence-killing-us-all.html]

Christians, C. G. 2019. Media Ethics and global Justice in the Digital Age. Cambridge: Cambridge University Press.

Cortese, F. A. B., 2014. 'The Maximally Distributed Intelligence Explosion'. Implementing Selves with Safe Motivational Systems and Motivational Systems and Self-Improvement: Proceedings of the AAAI Spring Symposium: 7–12.

Cotton-Barratt, O., Daniel, M., and A. Sandberg. 2020. 'Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why they all matter'. Global Policy: 1–12. doi: 10.1111/1758-5899.12786

Dafoe, A., 2018. 'AI Governance: A Research Agenda'. Oxford: Future of Humanity Institute. [https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf]

Danaher, J., 2015. 'Why AI Doomsayers are Like Sceptical Theists and Why it Matters'. Minds & Machines 25: 231–246.

Dehaene S., Lau H., and S. Kouider. 2017. 'What is consciousness, and could machines have it?'. *Science* 358(6362): 486-492. doi:10.1126/science.aan8871

Dignum, V. 2019. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer International Publishing.

Drexler, K. E., 2019. 'Reframing Superintelligence: Comprehensive AI Services as General Intelligence'. Technical Report #2019-1, Future of Humanity Institute, University of Oxford.

Future of Life Institute (FLI). 2017. 'Open letter to the United Nations Convention on Certain Conventional Weapons'. [https://futureoflife.org/autonomous-weapons-open-letter-2017/]

Geist, E., 2016. 'It's already too late to stop the AI arms race - We must manage it instead'. Bulletin of the Atomic Scientists 72: 318–321.

Geist, E. and A.J. Lohn, How Might Artificial Intelligence Affect the Risk of Nuclear War?. Santa Monica, CA: RAND Corporation, 2018. [https://www.rand.org/pubs/perspectives/PE296.html]

Good, I.J., 1965. 'Speculations Concerning the First Ultraintelligent Machine'. Advances in Computers, 6 (99): 31–83.

Groll, E., 2018. 'How AI could destabilize nuclear deterrence'. Foreign Policy, 24 April. [https://foreignpolicy.com/2018/04/24/how-ai-could-destabilize-nuclear-deterrence/]

Häggström, O., 2019. 'Challenges to the Omohundro–Bostrom framework for AI motivations'. Foresight 21 (1): 153–166.

Hawking, S., 2018. Brief Answers to the Big Questions, United States: Hodder & Stoughton Publishers.

Hume, D., 1739. A Treatise on Human Nature, Edited by D. F. Norton and M. J. Norton. 2000. 6th Edition. Oxford: Oxford University Press.

Huysmans, J., 2006. The Politics of Insecurity: Fear, Migration and Asylum in the EU, Oxford: Routledge Press.

Koch, C. and G. Tononi. 2017. 'Can We Quantify Machine Consciousness?'. IEEE Spectrum, 25 May 2017.

Koch, C., 2019. 'Proust among the Machines'. Scientific American 321(6): 46-49. doi:10.1038/scientificamerican1219-46

Kohs, G., (Director), Macdonald, A. (Producer), and A. Reich. (Producer). 2017. AlphaGo. Reel As Dirt.

Leslie, D., 2019. 'Raging Robots, Hapless Humans: The AI Dystopia'. Nature 574: 32-33.

Lim, D., 2019. Killer Robots and Human Dignity. Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, New York: 171–176.

Liu, H., Lauta, K. C., and M. M. Maas. 2018. 'Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research'. Futures 102: 6-19.

Loosemore, R., 2012. 'The fallacy of dumb superintelligence'. Institute for Ethics and Emerging Technologies Blog. [http://ieet.org/index.php/IEET/more/loosemore20121128]

Martin, B., 1982. 'Critique of nuclear extinction'. Journal of Peace Research. 19 (4): 287–300.

Matheny, J. G., 2007. Reducing the Risk of Human Extinction. Risk Analysis 27 (5): 1335–1344.

May, T., 2018. 'Would human extinction be a tragedy?'. New York Times: The Stone, 17 December. Accessed on 30 April 2020.

Metzinger, T., 2017. 'Benevolent Artificial Anti-Natalism (BAAN)'. The Edge. 8 July. [https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan]

Moynihan, T., 2020. 'Existential Risk and Human Extinction'. Futures 116: 102495.

Ó hÉigeartaigh, S., 2017. 'The State of Research in Existential Risk'. Proceedings of the First International Colloquium on Catastrophic and Existential Risk, B. John Garrick (Ed), Published by The B. John Garrick Institute for the Risk Sciences, UCLA Dec 2017: 37–52.

Omohundro, S. M., 2008. 'The Basic AI Drives'. Proceedings of the First Conference on Artificial General Intelligence, Amsterdam: 483–492.

Ord. T. 2020. The Precipice Existential Risk and the Future of Humanity, New York: Hachette Books.

Pinker, S., 2019. 'Tech prophecy and the underappreciated casual power of ideas'. In Possible Minds: Twenty Five Ways of Looking at AI, edited by J. Brockman, Chapter 10. Penguin Press.

Price, H. and K. Vold. 2018. 'Living with Artificial Intelligence'. Research Horizons, Issue 35. February 2018: 20-21.

Rees, M., 2003. Our Final Hour: A Scientist's Warning, New York: Basic Books.

Rickli, J., 2019. 'The destabilizing prospects of artificial intelligence for nuclear strategy, deterrence and stability'. In The Impact of Artificial Intelligence on Strategic Stability And Nuclear Risk, edited by V. Boulanin, 91-98. Solna, Sweden: SIPRI.

Robinson, J. 2016. 'U of T's Geoffrey Hinton: AI will eventually surpass the human brain but getting jokes ... that could take time'. U of T Magazine: News Online. [https://www.utoronto.ca/news/u-t-geoffrey-hinton-ai-will-eventually-surpass-human-brain-getting-jokes-could-take-time]

Roff, H. M., 2014. 'The Strategic Robot Problem: Lethal Autonomous Weapons in War'. Journal of Military Ethics, 13 (3): 211–227.

Rosert, E. and F. Sauer. 2019. 'Prohibiting Autonomous Weapons: Put Human Dignity First'. Global Policy 10: 370–375.

Russell, S., 2015. 'Take a stand on AI weapons'. Nature 521: 415–418.

Russell, S., 2019. Human Compatible: AI and the Problem of Control, USA: Penguin Random House.

Russell, S., and P. Norvig. 2010. Artificial Intelligence A Modern Approach
Third Edition. New Jersey: Prentice Hall.

Sagan, C., 1983. 'Nuclear war and climatic catastrophe: Some policy implications'. Foreign Affairs 62 (2): 257–292.

Sandberg, A. and N. Bostrom. 2008. 'Global Catastrophic Risks Survey'. Technical Report #2008-1, Future of Humanity Institute, Oxford University: 1–5.

Sauer, F., 2019. 'Military applications of artificial intelligence: Nuclear risk redux'. In The Impact of Artificial Intelligence on Strategic Stability And Nuclear Risk, edited by V. Boulanin, 84-90. Solna, Sweden: SIPRI.

Schulte, P., Alegret, L., Arenillas, I., Arz, J. A., Barton, P. J., Bown, P. R., et al. 2010. 'The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary'. Science 327 (5970): 1214–1218.

Searle, J., 1980, 'Minds, Brains and Programs'. Behavioral and Brain Sciences 3: 417–57

Shanahan, M., 2015. The Technological Singularity. Cambridge: MIT Press.

Sharkey, A., 2019. 'Autonomous weapons systems, killer robots and human dignity'. Ethics of Information Technology 21: 75–87.

Shevlin, H., Vold, K., Crosby, M., and M. Halina. 2019. 'The Limits of Machine Intelligence'. EMBO Report 20: e49177.

Shulman, C. 2012. 'Nuclear winter and human extinction: Q&A with Luke Oman'. Overcoming Bias Blog, 5 November. [http://www.overcomingbias.com/2012/11/nuclear-winter-and-human-extinction-qa-with-luke-oman.html]

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. 2016. 'Mastering the game of Go with deep neural networks and tree search'. Nature 529: 484-489.

Sofge, E., 2015. 'Bill Gates Fears A.I., But A.I. Researchers Know Better.' Popular Science, 30 January. [https://www.popsci.com/bill-gates-fears-ai-ai-researchers-know-better/]

Sotala, K., and L. Gloor. 2017. 'Superintelligence as a Cause or Cure for Risks of Astronomical Suffering'. Informatica 41: 389–400.

Sparrow R. 2007. 'Killer Robots,' *Journal of Applied Philosophy*. 24(1). 65.

Tegmark, M., 2014. 'Friendly artificial intelligence: the physics challenge'. Proceedings of the Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop: 87-89.

Tegmark, M., 2017. Life 3.0: Being Human in the Age of Artificial Intelligence, New York: Brockman Inc.

Torres, P., 2019. 'Existential risks: a philosophical analysis'. Inquiry, DOI: 10.1080/0020174X.2019.1658626

Turing, A., 1951. 'Intelligent Machinery, A Heretical Theory'. Reprinted in: Philosophia Mathematica 4 (3): 256–260. 1996.

Wakefield, J., 2015. 'Intelligent Machines: What does Facebook want with AI?'. BBC News, 15 September. [https://www.bbc.com/news/technology-34118481]

Weiner, N., 1960. 'Some Moral and Technical Consequences of Automation'. Science 131 (3410): 1355–1358.

Williams, C., 2015. 'AI guru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars'. The Register, 19 March. [https://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai/]

Zwetsloot, R. and A. Dafoe. 2019. 'Thinking about Risks From AI: Accidents, Misuse and Structure'. Lawfare Blog. 11 February. [https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure]