Artificial superintelligence and its limits: why AlphaZero cannot become a general agent

Abstract

An intelligent machine surpassing human intelligence across a wide set of skills has been proposed as a possible existential catastrophe (i.e., an event comparable in value to that of human extinction). Among those concerned about existential risk related to Artificial Intelligence (AI), it is common to assume that AI will not only be very intelligent, but also be a general agent (i.e., an agent capable of action in many different contexts).

This article explores the characteristics of machine agency, and what it would mean for a machine to become a general agent. In particular, it does so by articulating some important differences between belief and desire in the context of machine agency. One such difference is that while an agent can by itself acquire new beliefs through learning, desires need to be derived from preexisting desires or acquired with the help of an external influence. Such influence could be a human programmer or natural selection. We argue that to become a general agent, a machine needs *productive* desires, or desires that can direct behavior across multiple contexts. However, productive desires cannot *sui generis* be derived from non-productive desires. Thus, even though general agency in AI could in principle be created, it cannot be produced by an AI spontaneously through an endogenous process. In conclusion, we argue that a common AI scenario, where general agency suddenly emerges in a non-general agent AI, is not plausible.

1. The Intelligence explosion

A number of scholars, most prominently Alan Turing (1950) have argued that humanity will eventually create an Artificial Intelligence (AI) with greater intelligence than that of any human. Such a machine would have, it is argued, superior ability to improve itself and/or to create improved versions of itself. If this happens, the AI could not only improve its abilities to solve a wide range of problems, but also its ability to improve itself. The result of this iterative process, it is argued, would be a machine with intelligence far surpassing that of any human or human organization. In other words, it would be *superintelligent*. In this context, "intelligence" means "the ability to change the world in order to attain one's goals". Thus, a superintelligent machine could, if its motivations were not well aligned with the interests of humanity, pose an existential risk. Various research agendas have been suggested for managing the challenges that such machines would pose to society (Soares and Fallenstein 2017), (S. Russell, Dewey, and Tegmark 2015), (Amodei et al. 2016).

It should be noted that this account differs from some well-known science fiction narratives (e.g. *Terminator*). First, while intelligent machines are typically stereotyped as being cognitively rigid and uncreative, we have no reason to believe that a superintelligent AI would have such weaknesses. Second, while popular narratives of intelligent machines typically feature *robots* killing people, the possibilities of a superintelligent machine are not limited to robots. AI could, in theory, be disembodied and act through social manipulation of humans. Third, a truly superintelligent AI cannot simply be "switched off". As with any sufficiently powerful and intelligent person, AI could conceivably predict efforts at undermining its goals and take pre-emptive action. In short, the AI under consideration is both superintelligent and, contrary to existing AI, also a general agent, capable of action across many situations. (Bostrom 2014; Haggstrom 2016; Tegmark 2017; Yampolskiy 2015).

Robin Hanson and others have argued that general agency seems unlikely to be explicitly built into future AI (Hanson 2016). As a result, some AI scenarios in the literature describe general AI agents that appear without explicit efforts. Two possibilities are prevalent in these scenarios. Here, we will argue that the first scenario is implausible. We believe that the second scenario is unlikely, but this claim will not be defended here. We merely describe it to distinguish it from the first scenario.

1) Spontaneous emergence - A non-general machine agent becomes a general agent as a by-product of its success in improving its optimization power. Bostrom argues that "agent-like behavior in AI can "emerge spontaneously" (Bostrom 2014, p 153, 155). Such AI, Bostrom envisages, could if intelligent enough, start improving itself and in the process acquire general agency.[1]

2) Accidental emergence - AI researchers unwittingly create a general agent while believing it to be a non-general agent. As the AI's intelligence improves, it learns about new ways to attain its broadly defined goals. This is the "paperclip maximizer" scenario, where a general machine agent with the instruction to produce paperclips becomes more intelligent and becomes hostile as it tries to use limited resources to construct paperclips.

[1] "Instead of allowing agent-like purposive behavior to emerge spontaneously and haphazardly from the implementation of powerful search processes (including processes searching for internal workplans and processes directly searching for solutions meeting some user specified criterion), it may be better to create agents on purpose." (Bostrom 2014, p 155)

In this article we will explore the role of the term "agency" in the context of the first scenario. A better understanding of the properties of agency can enhance our understanding of the risk landscape and which risk mitigation strategies are most appropriate. We start by introducing a substrate neutral definition of agency, which is applicable regardless of the specific implementation details of the agent being analyzed. This notion follows and expands on Russell and Norvig (2016). We will then discuss the notion of generality of agency, and how it may be understood. We then introduce the belief/desire model of intentional action. This model, that has been widely discussed in analytical philosophy, has been used to analyze non-person agents, such as groups and non-sentient organisms. We will argue that to be a general agent, AI needs a certain kind of desires, that we will refer to as "productive" desires. The conclusion of this analysis is that a non-general machine agent cannot by itself become a general agent. This implies that scenario (1) cannot happen. Finally, we discuss some objections.

2. Agency and action

An agent, generally speaking, is a being that has the ability to *act*, and agency is to exercise that ability.[2] According to the standard theory of action, acts require the ability to represent the world ("beliefs"), and a set of motivations or instructions on what to achieve in the world ("desires"). According to a popular metaphor, beliefs are a map, and desires are the destinations. This is also known as the "belief/desire model for intentional action" and remains the most widely accepted model for understanding action and agency. According to this model, often called the Humean theory of action after the 18th century philosopher David Hume, mere beliefs cannot cause action; desires are also necessary (Peterson 2000, p 145). As illustrated by the above-mentioned metaphor, a better and more detailed map will facilitate getting to one's destination. But improving the map cannot by itself provide one with a destination.

Note that terms such as "belief" and "desire" describe states from a "subjective" (or first-person) point of view. If we were to observe a human brain, we would not see physical objects that correspond to beliefs and desires. Neither would we be able to clearly distinguish between believing and desiring when looking at an fMRI scan. Does this mean that these terms are artificial or not real? We will not settle this thorny issue here. However, we only need to assume that these terms refer to something real from *the perspective of an agent*. Thus, while it may be true that, for a neurosurgeon or a chipmaker, there is little that differs between an agent's beliefs and desires, this

[2] Note that we not exploring the important but distinct notion of *moral agency*. For an excellent discussion of moral agency in machines see (Gunkel and Bryson 2014).

does not mean that, for example, that there is no difference for the agent between *believing* that it will rain is and *wanting* it to rain.

We adopt an instrumentalist stance with regards to agency (Dennett 1987). According to this view, it is appropriate to attribute states such as "beliefs" and "desires" to an entity when doing so supports successful predictions of that entity's behavior. This notion is similar to what (Barandiaran, Paolo, and Rohde 2009) refer to as "minimal agency". On their view, an agent is an entity that is separate from its environment and that displays behavior to attain its goals. They suggest that organisms as simple as bacteria exhibit this minimal kind of agency. In other words, the kind of agency that is of interest here does not presuppose subjective experience or consciousness. Nor does it presuppose having a brain or a nervous system.[3]

An agent can be more or less general. We refer to less general agents as "specialized agents". A thermostat is a (very) specialized agent: it can represent the world through its thermometer, and it has a set of instructions to activate an AC if the thermometer registers a certain threshold value. By contrast, a dog is a general agent. It can represent the world with a variety of sensory organs and has desires that can direct its behavior in many different situations. General agency is, loosely speaking, the ability to act in a *diverse* set of situations.[4] As such, agency comes in degrees. Some machine agents are more general than a thermostat, yet less general than a dog.

Agency is sometimes described as an aspect of intelligence in the AI literature (Bostrom 2014; Yudkowsky 2013), in particular in reference to superintelligent AI. Yet, this is an unfortunate mistake. Agency is distinct from "intelligence". Intelligence, in this context, is the ability of an agent to change the world in accordance with its desires. If an agent has very specialized desires, it can be superintelligent with regards to those desires without being general. For example, AlphaZero has the goal to organize the pieces on a chessboard (according to the rules of chess) into a winning position. With regards to this goal, it is superintelligent in the sense that it is better than any human. By contrast, cats are very general but not superintelligent. We may falsely believe that to be superintelligent, a machine also needs to be a general agent. But this is because the term "intelligence", as applied to people, typically describes very general agents. For example, imagine that we observe a lab experiment where a person is instructed to collect apples from a tree. The

[3] We should also note that the notion of agency under consideration is more minimalistic than that proposed by (Floridi and Sanders 2004). In other words, this discussion is unrelated to current ongoing discussions on whether AI can become a *moral agent,* a *moral patient* or be *morally responsible.*

[4] A set is more diverse the greater the expected dissimilarity between a randomly sampled object in that set and the most similar object in that set (Gustafsson 2010).

person cannot reach the apples and fails to realize that a nearby ladder could be used for this purpose. We would typically conclude that this person is unintelligent, because the instruction was to reach the apples, and did not constrain the method by which the goal was to be attained.

In many situations, machines acting in our environment are often described as trying to attain some general aim, for example "clean a room". This is also the case in some of the literature on AI risk (Tegmark 2017, add p nr). But vacuum cleaner robots are not literally trying to clean a room. When we see that a machine fails to attain the aim that we ascribe to it (perhaps by getting stuck), we conclude, by analogy to how we evaluate the intelligence of a person in a similar situation, that the machine is unintelligent. But in many cases, machines have much more specific instructions than the ones we ascribe to them. And sometimes machines achieve those instructions very well, even when seeming stupid to us. For example, we may mistakenly think that a thermostat has the instruction to keep the temperature comfortable in a room. If so, we might think that the thermostat would be unintelligent if it failed to close the blinders, which would have prevented sunlight making the room uncomfortably warm. But in fact, the thermostat has no desire with respect to the temperature of the room. It only has a desire to activate the AC if the temperature, as registered by its thermometer, reaches a certain threshold. A machine with a specific instruction is superintelligent if it is better than any human at optimizing the world according to that instruction. The term "intelligence" has a lot of connotations that are misleading (authors 2018).

3. Agency and productive desires

What makes an agent more general? Intuitively, the notion of an agent's generality is related to the extent to which agency can be expressed. In turn, this will depend, according to the belief/desire model, on whether an agent has the appropriate desires. Metaphorically speaking, one can only intentionally travel to a location where one wishes to travel. First, a definition of "desire":

> For an agent to desire *p* is for the agent to be disposed to take whatever actions it believes are likely to bring about *p*.

Note that, as (Petersson 2000, p 80) argues, the belief/desire model of action does not presuppose that desires must be felt, experienced or associated with any sensation.

As humans can typically express agency (i.e., act) across many diverse situations, and discussions in philosophy often concerns humans, our existing language takes for granted that agency is

general, and that desires by nature can direct behavior across many contexts. We seem to lack the vocabulary for distinguishing between desires that are as specific as those in machines and those typically found in humans. Therefore, we wish to introduce a new notion to distinguish between desires which are conducive to general agency (for example, the desire to enjoy life) and those who are not (for example, the desire to organize chess pieces into a winning position on a digital chessboard according to a specific set of rules). We call this notion "productivity". An agent is general to the extent that its *desires* are *productive*.

We define "productivity" as follows:

> Productivity is a feature of a desire which specifies the extent to which it can direct behavior in a diverse set of situations.

Typically, a productive desire can direct action by being able to, in combination with appropriate beliefs, produce new and often situation-specific desires. For example, a rabbit has the productive desire to satisfy its hunger. In combination with the belief that some tasty salad is present, a rabbit's desire can produce the situation-specific desire to eat the salad. By contrast, some desires typically found in contemporary AI are not very productive. For example, the desire of AlphaZero to organize the pieces on a chessboard into a winning position is relatively unproductive, because it can only produce desires (and hence behavioral output) with regards to a chessboard. AlphaZero does not have the more productive desire to win chess matches. If it did have this desire, it would, given adequate beliefs about the world, try to convince people to play against it. Machines that have specific instructions to maximize the points in a game under a specific set of conditions, produce desires that further that aim, for example by learning a specific pattern of behavior. Desires can be less productive if they are constrained to a specified domain, such as a digital chessboard. But desires can also be less productive if they are highly specific. For example, a thermostat has a thermometer. When the thermometer registers a certain temperature, it conveys a signal to the thermostat. The thermostat has an instruction to output a certain signal to the AC if and only if receiving that incoming signal. This desire is so specific that it cannot produce any other desire, even while not being constrained to a digital domain.

The fact that machine agents never produce desires unrelated to their initial desires does not imply that their behavior is perfectly predictable. By contrast, machines sometimes surprise us, and when they do so, it may seem as an instance where they were "thinking outside the box" by acting outside of their instructions. But this not the case. Rather, machines do what they have been instructed to

do, which is not always what we expect them to do. This is true for AI as well as for simpler machines. For example, if a thermostat's thermometer is exposed to direct sunlight, the thermostat will register the increased temperature and activate the AC even when the average temperature of the room is below the target value.

A desire to make paper clips is very productive compared to the desires of AlphaZero. While this desire may seem like an unproductive desire, we should recognize that, if stated as "maximize the number of paper clips", the desire could direct behavior and/or produce new instrumental desires in many situations. For example, if an agent with the desire to make paper clips acquires the belief that "humans will try to prevent me from making paper clips", it could produce a new desire to eliminate the threat. This in turn could produce many different and complex desires.

4. Can specialized agents acquire productive desires?

We should note that the belief/desire model stipulates an important difference between belief and desire. One such difference is that beliefs and desires have different "directions of fit"(McNaughton 1988). A belief represents the world as being in a state of affairs such that it is true. Beliefs, some philosophers have argued, aim at the truth and so aim to fit the world (Smith 1987). A belief is satisfied when it fits the world. Thus, the direction of fit is "mind to world". By contrast, desires aim to change the world and so has a **world-to-mind** direction of fit. A desire, unlike a belief, doesn't depict the world as being in a certain state; rather it disposes an agent to change the world to a certain state. Desire is a state that is satisfied when the world fits it.

This difference matters because learning, or the acquisition of new beliefs and desires, requires some kind of reinforcement. Here, beliefs can (often) be directly confirmed by interacting with the world. If I believe that it rains, I can confirm or falsify that hypothesis by looking through the window. By contrast, the only way I can wish it to rain is if I have a desire that is somehow connected to rain. The world can reinforce my beliefs, but not my desires. My desires can only be reinforced from within, by my own desires.

In other words, since a desire can only be acquired from a set of pre-existing desires, an AI with a set of desires constrained to a specific domain cannot acquire desires relevant to other domains. For example, AlphaZero has a set of desires that pertain to the organizing of chess pieces. Making paper clips is not among those desires. Moreover, AlphaZero has no desire that can produce the desire to make paper clips, no matter how much it improves its ability to organize chess pieces. By contrast, if a hypothetical AI with the desire to produce paper clips thought that it could further its aims by defeating AlphaZero at chess, it would produce the desire to do so. In the terminology

introduced in the previous section, an AI with the productive desire to make paper clips can acquire the desire to play chess. But since AlphaZero's desire is unproductive, it cannot produce the desire to make paper clips.

If an agent can only acquire desires that are relevant for a particular domain or set of pre-existing desires, then it cannot acquire desires that are more productive than the desires it already has. Doing so would imply that it can acquire desires that are relevant for other things than its domain. For example, AlphaZero has the desire to organize a chessboard into a winning position. From this desire, it can produce the desire to move d2-d3. This new desire is less productive than the desire that produced it, since this new desire cannot produce the desire to organize the chessboard into a winning position.

If these claims are convincing, then there are interesting implications for the scenario described by Bostrom, where a specialized AI agent with a set of unproductive desires acquires new and more productive desires as it becomes better at processing information (Bostrom 2014). To become a more general agent, it needs to acquire more productive desires. But unproductive desires cannot produce more productive desires. Thus, a specialized AI agent will behave according to its desires, and as long as its desires are unproductive, it will remain specialized. This means that scenario 1, as described in the introduction, where general agency appears suddenly through an endogenous process in a non-general machine, cannot happen.

5. Objections

Latent general agency. In the AI-x-risk literature, there are two main scenarios regarding the unintended emergence of agent AI, the second being the creation of an unintelligent but general machine agent.

Consider the following scenario: *Iron Ore AI*, a seemingly specialized agent has a very productive desire; "produce pig iron from iron ore" but very few beliefs and quite limited intelligence. Since this machine is limited, its creators falsely believe it to be a specialized AI. However, Iron Ore AI can improve its intelligence, and as it does so, it acquires the belief that it can produce pig iron more effectively if it is even more intelligent. As it becomes even more intelligent and acquires more true beliefs about the world, it learns that there is a lot of iron ore under human settlements, and that the most effective way of extracting it is to remove those settlements. It also understands

that humans resent being forcibly resettled and starts planning measures to prevent humans from being able to pose a threat to its activities.

Iron Ore AI is in some respects similar to a dog, a very general agent limited only by its low intelligence. Creating a machine-like Iron Ore AI would be reckless, as suggested by this scenario. But creating such a sophisticated machine is not likely to be the result of a mistake. The idea that Iron Ore AI has any passing resemblance to existing AI rests on the fallacy of ascribing productivity to the desires of machines. Iron Ore AI is only likely to be created by someone who wants to create a general agent. However, a thorough case against this scenario is beyond the scope of this paper.

Self-preservation. If AI has a specific goal, wouldn't it also infer the goal of preserving itself given enough knowledge on how to attain its goals? In Life 3.0, Max Tegmark (2017) exemplifies this thesis with a game where the computer directs a character in a maze. The computer is rewarded when the character saves some sheep, and the character can be "killed" by a wolf. Tegmark notes that the computer learns that it cannot save the sheep if it is killed by the wolf, and thus acquires the desire to preserve its life. This way of describing desires is typical of the literature. In fact, the computer acquired a desire for self-preservation, nor does it aim to maximize the number of saved sheep. It has merely learned that one particular sequence of left/right/up/down moves maximizes its score. Suggesting that the computer has learned self-preservation would imply that it now will avoid other digital predators in the maze. This is not true. As we have argued, these desires can only be produced by a sufficiently productive desire. This machine does not have a desire to maximize an entity in the world. Rather, it aims to maximize its score by trying out sequences of left/right/up/down moves. The desire "maximize the number of paperclips in the world" is very different from the desire of the game described by Tegmark. Whereas the first instruction would not imply any desire to self-preservation, the second instruction might.

Natural selection. Agency in humans and other animals emerged as a result of the interaction of specialized agents with a changing environment and random mutations. This fact could be interpreted as a counterexample to the thesis advanced in this article. In natural evolution, the process that produced productive desires in animals took billions of years. Evolution by means of natural selection required an external force (selection pressures) and random variation. In this case, selection pressure is a kind of "programmer", an external force that directs the evolution of certain traits. This shows that the emergence of animal agency was not a result of an endogenous process,

but a result of a complex interaction of specialized agents, random chance and selection pressures. Thus, natural selection suggests one way in which an AI could become more general without *human* involvement. However, the process by which evolutionary computation creates algorithms is notoriously slow and inefficient unless guided by agents (AI researchers) to achieve certain traits. If general agents machine agents are possible, they are much more likely to emerge as a result of intelligent design than a result of natural selection.

Pain. Some specialized AI learn by negative reinforcement (a loss function). This form of negative reinforcement can be perhaps be described as pain. Yet, there are important differences between the functional role pain plays in animal cognition and in reinforcement learning in contemporary AI. An AI that is "punished" for losing at chess is only punished for losing at chess. The same AI would need to be reprogrammed with a new set of loss functions to play Tetris. By contrast, what causes pain in animals can be described as a very general set of instructions. When humans are subjected to pain, they not only learn to avoid the specific cause of the pain, they generalize across many similar causes. The "code" for what causes pain in an animal is thus much more general than the code for negative reinforcement of AlphaZero.

While "avoid pain" sounds like a very specific desire, it is actually very general. Typically, machines are suited with instructions to deal with specific threats to their structural integrity in specific ways. For example, what does "avoid" mean? A machine needs to be very general to understand what "avoid" means in a natural environment. We do not doubt that it may be possible to program a machine to experience pain as animals do. This would be a very general agent, like Iron Ore AI. But we are nowhere near any such machine.

Model agency/real-world agency.[5] Many existing AI agents are "inside model" (i.e., they have only beliefs and desires relevant to a specific model of the world). AlphaZero is one such agent. However, this is not true of some other machine agents (for example a Roomba, an autonomous vacuum cleaner). One objection to our argument is that desires constrained to models cannot be general, but desires about the real world are different. Even machine agents that are very specialized can, if they have desires that pertain to the real world, become general as they accumulate more beliefs about how to attain their goals. For example, an automated lawn mower may realize that sticks impede its functioning. This may in turn allow the machine to infer a desire to remove a tree that produces sticks.

[5] We are grateful to Linda Linsefors for this objection.

We believe that this objection implicitly assumes that machine desires about the real world are more productive than we have reason to believe. In the case of AlphaZero, it is obvious that its desires are specialized. But with machines that interact with the world, it is less obvious, and more tempting to anthropomorphize machine behavior. However, consider a machine like the Roomba. While its behavior may seem to suggest that it has productive desires, upon closer consideration, we find very specialized desires. For example: "random walk-angle chance upon bumping into an object". It is in fact quite remarkable, and the main idea behind roboticist Rodney Brooks AI design philosophy, of how much useful behavior it is possible to generate with very simple desires (Brooks 1990).

Finally, we have no reason to believe that there is a fundamental difference between "model-related" and "real-world" beliefs and desires in this context. Both play fundamentally the same role with regards to action.

6. Concluding remarks

We have argued that the belief/desire model of the concept of "agency" is useful in the analysis of machine agents. We have argued that the belief/desire model postulates some important differences between belief and desire. These difference matter for at least one of the scenarios discussed in the AI risk literature. The belief/desire model predicts that no specialized AI agent could become a general agent merely through an endogenous process of self-improvement.

This analysis is good news from an AI risk perspective. While we concede the possibility that a general AI agent is possible, our analysis suggests that possibilities of limiting the creation of such technology are realistic.[6] Rather than being complacent about AI risk, the conclusions from this analysis is that measures to monitor and guide the development of AI are both feasible and desirable.

References:

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *arXiv:1606.06565 [Cs]*, June. http://arxiv.org/abs/1606.06565.

[6] As Kaj Sotala (2018) and others have argued, there are multiple trajectories to superintelligent AI. This article has only explored one of them.

Barandiaran, Xabier E., Ezequiel A. Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action." *Adaptive Behaviour* 17: 367–86. https://doi.org/10.1177/1059712309343819.

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

Brooks, Rodney A. 1990. "Elephants Don't Play Chess." *Robotics and Autonomous Systems*, Designing Autonomous Agents, 6 (1): 3–15. https://doi.org/10.1016/S0921-8890(05)80025-9.

Dennett, Daniel. 1987. *The Intentional Stance*. MIT Press.

Floridi, Luciano, and J.W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3): 349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Gunkel, David J., and Joanna Bryson. 2014. "Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient." *Philosophy & Technology* 27 (1): 5–8. https://doi.org/10.1007/s13347-014-0151-1.

Gustafsson, Johan E. 2010. "Freedom of Choice and Expected Compromise." *Social Choice and Welfare* 35 (1): 65–79. https://doi.org/10.1007/s00355-009-0430-4.

Haggstrom, Olle. 2016. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford University Press.

Hanson, Robin. 2016. *The Age of Em: Work, Love and Life When Robots Rule the Earth*. 1 edition. Oxford: Oxford University Press.

McNaughton, David. 1988. *Moral Vision: An Introduction to Ethics*. Blackwell.

Petersson, Björn. 2000. *Belief & Desire the Standard Model of Intentional Action : Critique and Defence*. Björn Petersson, Dep. Of Philosophy, Kungshuset, Lundag\aard, Se-222 22 Lund,.

Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine* 36 (4): 105–14. https://doi.org/10.1609/aimag.v36i4.2577.

Russell, Stuart J., and Peter Norvig. 2016. *Artificial Intelligence : A Modern Approach*. Malaysia; Pearson Education Limited,. http://thuvienso.thanglong.edu.vn/handle/DHTL_123456789/4010.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind* 96 (381): 36–61.

Soares, Nate, and Benya Fallenstein. 2017. "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda." In *The Technological Singularity: Managing the Journey*, edited by Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong, 103–25. The Frontiers Collection. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-54033-6_5.

Sotala, Kaj. 2018. "Disjunctive Scenarios of Catastrophic AI Risk." Artificial Intelligence Safety and Security. July 27, 2018. https://doi.org/10.1201/9781351251389-22.

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

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (October): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Yampolskiy, Roman V. 2015. *Artificial Superintelligence: A Futuristic Approach*. 2015 edition. Boca Raton: Chapman and Hall/CRC.

Yudkowsky, Eliezer. 2013. "Intelligence Explosion Microeconomics." Technical Report 2013–1. Machine Intelligence Research Institute.