On the Logical Impossibility of Solving the Control Problem

Final Draft

Caleb Rudnick

November 5, 2019

Abstract

In the philosophy of artificial intelligence (AI) we are often warned of machines built with the best possible intentions, killing everyone on the planet and in some cases, everything in our light cone. At the same time, however, we are also told of the utopian worlds that could be created with just a single superintelligent mind. If we're ever to live in that utopia (or just avoid dystopia) it's necessary we solve the control problem. The control problem asks how humans would be able to control an AI arbitrarily. Nick Bostrom and other AI researchers have proposed different theoretical solutions to the control problem. In this paper, I will not look at the empirical question of how to solve the control problem. Instead, I will ask if we can solve it at all, a critical assumption most AI researchers have made is that we can. I propose, in fact, that we have a priori grounds for believing it is logically impossible to solve the control problem, since all superintelligent minds are, by definition, uncontrollable.

Introduction

The *problem* of the control problem is put best (and perhaps discussed most) by Nick Bostrom (2014:127) as the question of how "the sponsor of a project that aims to develop superintelligence ensure that the project [...] produces a superintelligence that would realize the sponsor's goals?" This definition is technically correct but makes it sound like a solution to the control problem is something that would be nice to have. When really, solving the control problem is a necessary step that needs to be taken, if we're ever going to develop an AI that values human life, takes less intelligent beings into consideration or just respects the planet. In short, solving the control problem is necessary but not sufficient if we want to mitigate the existential risk of an artificial superintelligence (ASI), since, a primary goal of (almost) any AI project on earth will be that it doesn't end our species.

Before continuing it's worth making precise what I mean by *artificial intelligence*, *superintelligence*, and *artificial superintelligence*. Today, each of these words can often be used interchangeably and for different purposes. In the philosophy of AI, not all artificial intelligences are superintelligent and not all superintelligences are artificial. Even a simple definition of artificial intelligence is not easy. Speaking broadly, AI is often viewed as "[making] computer systems (of various kinds) do what minds can do" (Boden, 2016). For our purposes, this definition of AI is sufficient. Yet, this definition only covers what the field of AI deals with as a whole, there is still a question of how *strong* a given AI system is. For instance, take *AlphaGo*, the computer that beat the best human at the board game *Go* (Bringsjord and Govindarajulu, 2018:§3.4), while this is certainly an example of a computer performing tasks that require human intelligence, it is only performing one of those tasks at a superhuman level, namely gameplaying. This is a weak or narrow AI because it is not performing, at even close to human levels, the full suite of tasks an average child could perform, like talking, understanding languages and recognising images (Bostrom, 2014:31). Narrow AI is probably the kind we hear most about in the news.

Importantly, narrow AI is not the kind I will be talking about in this paper. Instead, the control problem applies only to artificial superintelligences. This is an AI which "greatly [outperforms] the best current human minds across many very general cognitive domains" (Bostrom, 2014:52). These cognitive domains are of the kind we talked about earlier but also include things like "strategic planning" and "psychological modelling" (Bostrom, 2014:94).

However, just because something is superintelligent, does not mean it is artificial. Bostrom (2014:36) shows that, in principle, selective breeding could create a biological superintelligence, and George Dyson (2019) believes that superintelligence is only achievable through analogue systems. In this paper, however, ASIs will be our only concern. Finally, an AI which doesn't exceed but just matches human levels at a variety of tasks that require intelligence, we can call an artificial general intelligence (AGI). An AGI, then, is an AI that is equally intelligent to a human in many domains.

Based on the last paragraph, you might think the control problem is a moot point altogether. That is, if an ASI can perform tasks that require human intelligence, at a superhuman level, then surely one of the tasks it can perform, at a superhuman level, is ethical reasoning. In other words, if we do build an ASI, then it'll be smart enough to know that it shouldn't kill humans or harm the planet. Besides for the fact that the control problem is about controlling ASIs arbitrarily and not just in ways which benefit us, it's also not true that ASIs will necessarily have *superior* ethical reasoning to us, in the sense that what is objectively superhuman ethical reasoning might not be in our interests (Bostrom, 2003:277). After assessing all its options, an ASI could conclude that the most ethical thing to do is release a cloud of gas which blocks any feelings besides happiness in humans. This is an example of a superintelligence making a moral judgement that we would probably disagree with, if it had taken our interests into account. So just because an AI is superintelligent does not mean we should expect it to automatically value what we value.

Nick Bostrom and Eliezer Yudkowsky (2014:332) were right when they said that one of the most central issues in AI is figuring out "how to build an AI which, when it executes, becomes more ethical than you." But this question *is* ultimately a moot point if we can't solve the control problem, either in time or at all. I think that we have good a priori reasons for believing that the control problem is, in fact, unsolvable at all. My task for this paper will be to show why this is the case and why there is therefore no possible world where humans succeed in controlling an ASI. In section 1, I will give my argument for why there is no possible world where humans solve the control problem. In section 2, I will look at what some of the strongest objections to this argument are. Finally, in section 3, I will respond to these objections and attempt to show that they are not fatal to my argument.

1. The Problem with the Control Problem

In the philosophy of AI and empirical AI research, it's generally accepted that the control problem is solvable. Most current answers to the question 'can we, in principle, solve the control problem' come in two flavours: 'yes and this is how', with the other just being a plain 'no.' There is no camp that only gives logical or metaphysical reasons to think that the control problem is solvable, without also mentioning how to solve it. Let's briefly look at what reasons each of these camps give for their positions.

The 'yes and this is how' contingent take up a majority of researchers and philosophers. At the forefront of the control problem in philosophy and AI research, Bostrom falls into this camp. In his book, Bostrom (2014:127) lays out exactly what the control problem is and why it is such a hard problem to solve. While he doesn't give us reasons to think that humans can, in principle, solve the control problem, he does give us a variety of different methods to potentially solve the control problem, in the actual world. In a sense, Bostrom answers 'yes' implicitly, to the question of whether its logically possible to solve the control problem. If there is a method to control an ASI in the actual world, then that implies it's possible for us to solve the problem. Any solution to the control problem would come in the form of some kind of method to do the controlling. Bostrom (2014:143) gives several possible solutions to the control problem. Take the boxing solution to the control problem, where Bostrom proposes simply that we lock all ASIs in metaphorical boxes to stop them from doing things we don't want them to do. Posing this as a solution to the control problem in the actual world means that (1) the control problem is possibly solvable, while the solution might not work in the actual world it could work in some other possible world. From this we can infer (2), the control problem is in principle possible to solve by humans, because there is at least one world where humans solve it.

Eliezer Yudkowsky has, in conjunction with Bostrom, given a crucial reason for believing the 'yes and this is how' camp. This is that, as self-modifying systems, we can give an AI some instruction early on, say before it is superintelligent, such that the AI carries this instruction through to its superintelligent future self. In this way, even once an AI becomes an ASI, it will have been effectively controlled by humans (Yudkowsky and Bostrom, 2014:330). If this is true, then the control problem is solvable because it doesn't involve controlling an ASI but rather an AGI at an early stage,

which is logically possible. However, I don't think this line of reasoning works and in section 3, I will speak more about why this is not a genuine case of control.

In contrast, there are also some who take the position that the control problem *is* logically impossible to solve. As Yudkowsky and Bostrom (2014:330) note, Ray Kurzweil (2005) is one of these people. For Kurzweil "intelligence is inherently impossible to control" so there are no methods that would be able to control an AGI yet alone an ASI.

Another person I would include in this camp is Daniel Dennet. Though Dennett (2019:52) does not call ASIs uncontrollable, as explicitly as Kurzweil, he does imply as much. Saying that, if we were to imagine an ASI or perhaps even an AGI signing a contract with a human, there is no "penalty for promise breaking" we could impose on the AI if it were in breach of contract (Dennett, 2019:52). This is due to the nature of both AGIs and ASIs being digital, they could back up their minds to many locations such that there is no way we could conceivably punish them. We can't put them in prison, and we can't kill them (assuming they are advanced enough). They are "like superman" (Dennett, 2019:52). This means, we could only imprison an AI, if the AI *wanted* to go to prison. The question of how we could make an AI want something is identical with the control problem. Dennett (2019:52) says that a solution to which is "systematically difficult [to make fool proof] given the presumed cunning and self-knowledge of [an] AI."

George Dyson does not believe superintelligence will be digital. But it does seem likely that he would still agree with Dennet that the control problem is impossible to solve. He says, almost as explicitly as Kurzweil, that we can't control intelligence. Dyson identifies the problem with controlling intelligence in a *law* of AI which says: "any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand" (Dyson, 2019:38). Dyson is quick to show we can build things we don't understand, even so, if the AI we're seeking to control is so complicated we don't understand it, then it seems to follow that controlling it will be impossible and as such "our relationship with [it] will always be a matter of faith" (Dyson, 2019:38).

However, none of these accounts are sufficient to either, believe that we can, or that we cannot, control ASIs. Kurzweil and Dyson are correct that

controlling an ASI is impossible, but they have not spelled out why this is so. I think there are a priori grounds for believing that we can't control ASIs. My argument may even be going too far for those who, like Kurzweil, believe controlling ASIs *is* impossible, since they could really be using *impossible* in the way we use it in everyday speech and not in the logical sense. However, I do view the control problem as being impossible in the logical sense. Where logical impossibility is distinct from everyday or physical impossibility, in that, everyday physical impossibility expresses that something can't be true due to the nature of our world, while the logical sense of impossibility expresses that something is not true in all possible worlds. In the logical sense then, the impossibility of solving the control problem follows from how we define an ASI. Though there are plenty of definitions for 'ASI', any definition that captures the essential features of an ASI will work. At the start I took Bostrom's definition to be correct, that an ASI is an AI which "greatly [outperforms] the best current human minds across many very general cognitive domains" (Bostrom, 2014:52). But Irving John Good's definition also seems to capture the same essential truth about ASIs that Bostrom's does, when Good says (as cited in Bostrom, 2014:4) "an ultraintelligent machine [can] be defined as a machine that can far surpass all the intellectual activities of any man however clever." This common truth present in both definitions, is that ASIs are *essentially* more intelligent than humans. From this and using the following line of reasoning, I believe we can say that there is no possible world where the control problem is solved by humans:

1. A truly superintelligent artificial intelligence is essentially more intelligent than all humans

2. Let x be a truly superintelligent artificial intelligence

3. A solution to the control problem will let the solver in some way limit what x can do

4. If the control problem is solved by humans, then at least one human will have a way of limiting what x can do

5. If y limits what x can do, then y has a greater than or equal intelligence to x

6. Assume that humans have solved the control problem, then by

(4) this means at least one human has limited what x can do

7. Further, by (5) this means at least one human has greater than or equal intelligence to x

8. Yet, according to (1, 2, 5) if humans have a greater than or equal intelligence to x, then x is not truly superintelligent, since there is at least one human which is at least as intelligent as x

9. (8) and (2) being true is a contradiction

10. By *reductio ad absurdum* (6, 2, 8) the assumption that humans solve the control problem leads to a contradiction, so it must be logically impossible for humans to solve the control problem

This argument relies on definitions of the control problem, ASI, and their relation to humans, so there isn't an investigation in the world we would need to conduct, to verify whether it's sound. I believe there is a priori support for my argument that it is not logically possible for humans to solve the control problem. Though just because the argument is based on definitions doesn't mean everyone will agree that each premise is true. However, I believe there are good reasons to think they are. Importantly, not all the premises need the same justifications, some merely logically follow from others. Premise (1), that ASIs are essentially more intelligent than all humans, does seem somewhat controversial. Likewise, the claim in premise (5) is also not obviously true when it says that, if you can limit what a truly superintelligent AI can do, then you must have as much or more intelligence than it. Yet, I believe there are good reasons to think both (1) and (5) are true. If you accept both (1) and (5), I don't believe there is any way to resist the conclusion that the control problem is logically impossible for humans to solve.

The question of what makes a given property an *essential* property instead of an *accidental* property is, by itself, an entire area of philosophy. As such, if we want to find out what properties are essential to being a human or a toaster, instead of those that are merely accidental, our answers might change depending on what view of essential properties we take to be true. Pretheoretically, when we talk about essential and accidental properties we understand them in the sense that an essential property of a thing is "a property that it must have" while an accidental property is one which

it only "happens to have but that it could lack" (Robertson and Atkins, 2018). Further, Robertson and Atkins (2018) show that the specific way we characterise the *must* and *could* of our pretheoretical understanding, is where our answers to what makes something an essential property might diverge, though, this does not mean that we will get radically different answers depending on the account of essential properties we take to be correct. There must still be certain properties that come out as essential and others that come out as accidental, no matter the account (Robertson and Atkins, 2018). A common way of understanding what essential properties are is the modal account. Where "P is an essential property of an object o just in case o has P in all possible worlds, whereas P is an accidental property of an object o just in case o has P but there is a possible world in which o lacks P" (Robertson and Atkins, 2018:§1). Though the modal account of essential properties is common, that does not mean it's without any flaws. However, in our case, using the modal account of essential properties to understand whether ASIs are essentially more intelligent than humans, will be sufficient, since this should be an uncontroversial essential property that any correct account will call essential, in the same way that any correct account will call being human an essential property of Socrates.

With this modal account of essential properties in place, why should we think that 'being more intelligent than all humans' is something an ASI will have in all possible worlds? This comes from the way we characterise what an ASI would be like. Fundamentally, our definition of an ASI was an AI which cognitively outperforms human minds. Based just on that definition, it seems we would have to say there is no possible world where an ASI could exist without cognitively outperforming all humans, since that thing would no longer be an ASI if there were human minds which outperformed it. In the same way, there is no possible world where Socrates is not human, because the thing we call Socrates could not be Socrates, if it weren't human. Being more intelligent than all humans is simply a property an ASI could not lack.

Yet we can still look at why we should define an ASI in this way. Bostrom (2014:53) has examined the kinds of "superpowers" we would expect an ASI to have, and from these it does seem reasonable to conclude that an ASI is something which would cognitively outperform humans. Take just the speed and quality superpowers that Bostrom (2014:53) mentions. Having the superpower of speed would allow that system to do the same thinking a human could but "much faster" (Bostrom, 2014:53). This means, if we

just took an average adult human and gave them the "multiple orders of magnitude" speed increase Bostrom (2014:53) talks about, we could end up with a human that thinks 10^4 times as fast as a typical person. Such a speed increase would let that human not only perform cognitive tasks, like moving chess pieces to the best location on a board, or writing essays, at superhuman speeds, but it would also change the way that human perceives time. You could think about each chess move for subjective days or weeks while your opponent has mere minutes to perform the same thoughts (Bostrom, 2014:43). Along with this, the quality of thinking in a superintelligence is its own superpower. Where a superintelligence would think not faster but *better* than a human, in the same way that a human thinks better than an ant. Something with an ant's quality of intelligence simply cannot perform tasks like playing chess. To an agent that has a quality of intelligence superpower, there might be tasks it can perform that will seem as foreign to a human as chess does to an ant (Bostrom, 2014:57).

If superintelligence has just one of these superpowers, then I think it's safe to say that an artificial superintelligence will cognitively outperform humans and thereby be more intelligent than all humans.

The second crucial premise of the argument is (5) which says that if you manage to limit what an ASI can do then you must be at least as intelligent as it. By this, I don't mean it as a general rule that when you limit what something can do, then you must be at least as intelligent as it. There are several examples where this isn't the case, perhaps the most famous being the parasitic fungus *Ophiocordyceps*, which can greatly limit and "control ants' behaviour" (Harmon, 2012). Such that, if an ant becomes infected, the fungus will make the ant bite down on a leaf at the high point of a tree where the fungus will eventually grow out of its body and infect even more ants (Harmon, 2012). In this case, the fungus is less intelligent than the ant, in terms of the quality of its intelligence, but is still able to control and limit it in a significant way. What's more, Fredericksen et al. (2017) show that in an infected ant, the fungus is not found in the ants' brain but rather its muscles, which suggests the fungus controls the ants through, what Bostrom (2014:138) would call, a "capability control" method instead of a "motivational selection" method. This is because the fungus appears to work not by making an infected ant *want* to kill itself, but rather by directly affecting the ants' muscles such that the ant kills itself.

Since premise (5) only deals with ASIs, problems like the zombie ant do

not affect its truth. By this I mean, strictly speaking, the zombie ant is not a counter example to premise (5) which says: 'If y limits what x [a truly superintelligent artificial intelligence] can do, then y has a greater than or equal intelligence to x [a truly superintelligent artificial intelligence].' In the example of the zombie ant, the less intelligent fungus *is* genuinely limiting what a more intelligent ant can do but it is not limiting what a superintelligence can do. For the zombie ant to be a problem for premise (5) we would need a case of the fungus limiting what a truly superintelligent artificial intelligence can do. The way premise (5) is currently formulated is a weak claim and the case of the zombie ant would be a problem, only if I had formulated premise (5) more strongly, as the claim that: 'If y limits what x [a more intelligent agent] can do, then y has a greater than or equal intelligence to x [that agent].' This is because the stronger formulation of premise (5) is really making the universal claim that for all agents, it isn't the case that a less intelligent agent could limit a more intelligent agent. The case of the zombie ant then, disproves that stronger universal claim because by knowing that a fungus can limit an ant, we effectively know that there exists a less intelligent agent that can limit a more intelligent agent. The universal claim that, all less intelligent agents can never limit all more intelligent agents, is far too strong for me to commit to. Zombie ants are, themselves, a good reason to specifically formulate premise (5) more weakly, such that it doesn't make a claim about all more intelligent agents. Of course, even the weaker version of premise (5), which deals only with ASIs, is still a universal claim. This is because it talks about *all* agents which could limit an ASI. Yet, this is still a weaker claim since, for the stronger claim, just one thing existing that is less intelligent, which limits just one thing that is more intelligent, would be enough to make that claim false. However, for the weaker claim, we would need there to exist one thing that is less intelligent, which limits specifically a superintelligence, to make the claim false.

Even though the zombie ant shouldn't technically be able to make premise (5) false, we can still look at where the zombie ant puts pressure on premise (5). As I see it, the zombie ant forces us to ask why one should formulate premise (5) to only include ASIs, that is, why ASIs are unique in requiring intelligence to limit them, even though this doesn't apply to all intelligences? I think a true answer to this question is unknown, but I do believe there is genuinely something different about ASIs which gives them different properties to other intelligences, like ants and fungi. The most obvious difference is just the raw intelligence of an ASI compared to ants, fungi and even humans. Just like we seem to find consciousness in creatures with enough

brain power, so too it also might be, when we reach a further threshold of intelligence, limitations become impossible without sufficient intelligence.

The *Ophiocordyceps* fungus has managed to implement a significant method of capability control which directly forces some ants to do what the fungus wants, is there an analogous way that humans could work like the fungus or create a 'fungus' such that we directly force an ASI to do what we want? I think the answer is no. Like other capability control methods, that is, methods of control which limit what an ASI can do rather than what it wants to do, even we can think of ways an ASI will be able to escape them (Bostrom, 2014:146). No doubt, an ASI will think of ways to escape that we can't hope to pre-empt. The closest forms of capability control to a human made fungus, that Bostrom (2014:146) mentions, are boxing methods. This is where we attempt to physically contain an ASI such that it can't affect the outside world, for example, by unplugging the internet and building the ASI at the bottom of the ocean. Like the fungus, this method of control attempts to make an ASI directly do something we want it to do, namely, not killing people. However, even if at the bottom of the ocean we only communicated with this ASI via a text terminal, way up on the surface, it's almost certainly still going to be able to escape. This is because such a situation replicates exactly an experiment done by Yudkowsky (2002) where one person acts as a gatekeeper and another as an ASI in a box wanting to get out, they only communicate through text and the gatekeeper gets $10 if they don't let the ASI out. Yudkowsky has acted as an ASI in this situation twice and in both cases the gatekeepers let him, acting as the ASI, out of the box. Yudkowsky hasn't explained how he got the gatekeepers to let him out but even so we know that a human level of intelligence has been able to escape this kind of box and crucially, just that he escaped is enough to know an ASI might also be able to, since, "what a human can do, a superintelligence can do too" (Bostrom, 2014:283). The way fungus creates zombie ants is a kind of boxing method and yet ants can't escape it, while we can expect that superintelligences will be able to. When you get to a certain level of intelligence it seems there is less wiggle room for something like a 'fungus' to work as a control method. I think this is one good reason to think that there is genuinely something different about ASIs which links intelligence to limitations, in a way that other intelligences aren't.

Another category of control methods that Bostrom (2014:138) identifies are motivational selection methods, where we limit what an ASI wants to do. That is, giving an ASI rules so that it doesn't *want* to do what we

don't want it to do. Motivational selection methods come with their own problems and in section 3, I will look at these. Yet, I think with what we have seen so far, it's safe to say that limiting what an ASI can do is a difference in kind to limiting what an ant can do. There is a way in which intelligence is uniquely tied to applying limitations to ASIs that it isn't in other cases or at least isn't as essential in other cases. Specifically, that successfully applying limitations to an ASI depends on you being at least as intelligent, if not more intelligent than the ASI. To me, more raw intelligence is the only way to get around the problems with limiting an ASI that pre-empts, or works faster than you in all domains. However, this doesn't mean it's impossible to control an ASI, just that it's impossible for less intelligent minds like humans to control an ASI. This leaves open the possibility some super-superintelligence could still limit the abilities of an ASI. If such a super-superintelligence existed and wanted to limit an ordinary ASI, then it would be able to pre-empt any defences from the ordinary ASI in exactly the same way an ordinary ASI could pre-empt and avoid our best attempts at limitations.

ASIs will be superintelligent across many domains in the same way that today's best gameplaying AIs are superintelligent across only one domain. If you are playing a fair game of chess or Go with *AlphaZero*, it is not physically possible for you to win. It might be physically possible for you to beat *AlphaZero* at chess if you were not playing fairly and ignored some rules of chess and, for instance, were allowed to take days to prepare each move while *AlphaZero* were only allowed minutes. Or it might also be physically possible for you to win against *AlphaZero* if you had an even more powerful computer helping you make moves. Yet, barring these 'cheats', a game of chess between you and *AlphaZero* depends on just your intelligence versus its intelligence at chess and gameplaying. Analogously, if successfully applying control methods to an ASI is like a game of chess with a superintelligent narrow AI, then the rules of such a game of control would be like chess, only, instead of applying to a gameboard, they would apply to all the domains of the world. The notion of a fair game in this analogy corresponds to not breaking the physical laws of the universe, which is to say that, effectively, if you are both abiding by the same laws of the universe, a game of 'can I control an ASI', will again depend on your intelligence versus its intelligence, just like in the chess match. In the real world, the only rules that can be broken are physical laws. Assuming neither you nor the ASI can do this, all the game comes down to is just what the game came down to when it was a game of chess, intelligence versus intelligence. Only now the game is

intelligence-generally versus intelligence-generally, instead of intelligence-in-one domain versus intelligence-in-one-domain.

From this it follows that if a narrow AI can easily pre-empt our moves or make moves no one can predict in the first place, when it comes to chess, then we should expect a generally superintelligent AI to do the same sorts of things, only in real life and not just on a chess board. Whatever approach to control we take, be it surrounding an ASI in a faraday cage that lives in a bunker under the ocean, or attempting to engineer an ASI that will not want to hurt us, we will be making moves in a game all the same, though the game is much bigger and less will defined, if an ASI is truly superintelligent, then it will be able to shrug off those moves, which are our best attempts at control, in the same way that narrow superintelligences shrug of our moves in chess, which are our best attempts at beating them at chess. The only way we could ever hope to beat *AlphaZero* at chess fairly, is simply by being as intelligent or more intelligent than it. Likewise, I think it follows, that the only way to control an ASI is by being as intelligent or more intelligent than it. As was the case when *AlphaZero* beat its ancestor computer *AlphaGo* at chess, where it "achieved within 24 hours a superhuman level of play in the games of chess and [...] Go, and convincingly defeated a world-champion [chess and Go] program in each case" (Silver et al., 2017:1).

2. Problems with the Argument

There's a reason Bostrom (2014:255) said solving the control problem is "philosophy with a deadline" and that Yudkowsky (2001:7) said we have to "get it right the first time" and make zero errors, if we want to create a friendly AI. This is that both Bostrom and Yudkowsky want to leave open the possibility that we can solve the control problem and in so doing create an ASI that is aligned with our interests. Why they think this possibility exists is one of the strongest objections to the argument in section 1.

Motivation

In the previous section we looked at why it's not logically possible to control an ASI. However, what Bostrom and Yudkowsky will point out is that while it's perhaps logically impossible to limit what an ASI can do, it's not logically impossible to limit what an AGI can do, since AGIs are AIs that are

as generally intelligent as a human. If that's true, it might mean we could control an AGI in such a way that the AGI itself passes the control method on to any AIs the AGI constructs, or itself, if the AGI self modifies its mind to become superintelligent. In other words, if we came across some alien ASI that is already fully superintelligent, I'm sure Bostrom and Yudkowsky would take it as a trivial point that it would be impossible to control. However, if we make the ASI that we are seeking to control, then there is a chance we might be able to control it. Formally, this denies premise (5) of the argument in section 1, which said that if you limit an ASI, then you must be at least as intelligent as it. This is because, if it were true that you could limit an AGI and that limitation would pass on to a future ASI, we would have an example of something less intelligent (humans) limiting an ASI, so the premise would be false.

This denial of premise (5) comes from ideas about how we are likely to construct ASIs. Creating a fully superintelligent ASI from scratch, while perhaps logically possible, is not something we are likely to see in the actual world. Instead, what is more likely to happen is that we build what Yudkowsky (2007:479) calls a "seed AI." This is an AI which is capable of "self-understanding, self-modification, and recursive self-improvement"(Yudkowsky, 2007:479). At the start, this seed AI need not even be an AGI, that is, not even as generally intelligent as a human. Even so, using its self-modification abilities, the seed AI could change its own programming to become generally intelligent like a human is. From this point the seed AI could use the capabilities it has gained from being as intelligent as a human, to further self-modify to the point where it is superintelligent, creating an "intelligence explosion" (Bostrom, 2014:29). This is one of the more likely paths to superintelligence Bostrom (2014:22) has identified among others. The possibility of us realising this path means there might be a way for us to limit or incentivise what a future ASI can do, when it is at the level of an AGI or lower. Though this could depend on how long it takes for the AGI to self-modify into an ASI. If it is the "fast takeoff" Bostrom (2014:64) identifies, the seed AI could go from being an AGI to an ASI in less than a day. Yet, whether it is a fast or slow takeoff does not matter for our purposes, since either will open up the logical possibility of controlling what an ASI can do, before it becomes superintelligent and thereby prove that premise (5) is not correct.

One way of representing a motivational control method is as a *goal*. If an agent implements the goal successfully, then an AI will *want* to pursue it. Even assuming we do manage to implement a goal that an AI wants

to follow when it is at the level of an AGI, why should we think that it will carry through that goal to its future ASI self? We have good reason to think that if AGIs have minds that are anything like our own and express their intelligence in any way that is similar to ours, then they will be driven to think rationally, just as we are (Bringsjord and Govindarajulu, 2018:§2). In fact Omohundro (2008:485) identifies rationality as one of the basic drives an AGI or ASI will have. Since it would be hard to become superintelligent, if you were driven to think irrationally. One could imagine a chess computer that thinks irrationally attempting to lose as many games as possible instead of win as many as possible. This comes down to how you define rationality, but if we use a rational choice theory understanding of rationality, as Omohundro (2008:485) does, then its rational to take some action when it nets you more utility or expected utility than other actions. In this way, a rational chess computer that considered whether it should lose games instead of win them, would see losing as a negative expected utility and not take actions that lead to losing.

However we define the rationality an AI will have, what's important is the goal we decide to implement. Yudkowsky (2001:100) suggests that given the *right* goal, to the extent that an AGI is thinking rationally, it would be *irrational* for the AI not to persist that goal as the AI modifies itself, and so it would not "*want* to modify the goal" (Yudkowsky, 2001:100). This means that such a goal would have "to get the AI to see undesirable modifications as undesirable" (Yudkowsky, 2001:222). Such a goal might be something like a kind of axiology, which is "a certain theory of what things are good and how good they are" (Schroeder, 2016). Consider we give an AI a certain theory of good that identifies what is *good* with human happiness. If at some future point the AI becomes more intelligent and could potentially modify its axiology to something else other than human happiness, it doesn't seem likely the AI will commit that change since, according to its present axiology, this new axiology would be a negative utility. It would be like a chess computer considering whether it should change its goal from winning games to losing games. According to its current goal, the new goal would net only negative expected utility, if it were implemented. So, the chess computer would not pursue this new goal. This is all assuming there were a self-modifying chess computer. Yet, this does seem to follow for self-modifying AIs, if we look at the clear example from Yudkowsky and Bostrom (2014:330), which shows that we could expect this kind of behaviour from humans too. Gandhi had a value theory which made him not *want* to kill people. Were he presented with a magic button that could make him *not want* to *want* to be

peaceful, under his current axiology, pushing that button would be rather unattractive. So, in our case, where the prospect of a self-modifying AI being presented with an axiology changing button is a real possibility, we should expect the AI *not* to *want* to push the button. This is because, if the button were to change the AI's axiology as drastically as it did in the Gandhi example, then its *present* axiology would make pressing the button look like an unattractive option with a negative utility.

If we are the ones that implement such a goal or axiology in an AI, and that AI does indeed persist the goal from when it was only generally intelligent until it becomes superintelligent, then we have a clear example of something less intelligent controlling an ASI.

Consent

Along these lines, another strong objection to my argument is simply that an ASI could consent to us limiting it, which would likewise make premise (5) false that limiting an ASI is something you have to be at least as intelligent as it, to do. This applies equally well to an ASI we arbitrarily find in front of us, as it does to an ASI that we create through a seed AI. Let's take the second case first. Though we don't yet know how to build an ASI or AGI, if we build either of these, then it is likely that creating the AI will involve some form of programming. Though other paths to superintelligence mentioned by Bostrom (2014:22) include organic superintelligences, where instead of constructing a seed AI from scratch, we instead emulate a human mind on fast hardware which gives it access to more intelligence than it would otherwise have. However, if we take it that we are dealing with a superintelligence that humans create from a seed AI, then it seems inevitable this will have to involve some kind of computer programming. If this is true, and the seed AI does evolve into an AGI and then (eventually) an ASI, what will the initial code that contained our best attempts at solving the control problem look like to the AI? Though we can't know *what-it-is-like* for the AI to experience that code, Yudkowsky thinks it might be something like sensory perception. He writes:

> For a self-modifying AI with causal validity semantics, the presence of a particular line of code is equivalent to the historical fact that, at some point, a human wrote that piece of code. If the historical fact is

> not binding, then neither is the code itself. The human-written code is simply *sensory information* about what code the humans think should be written (Yudkowsky, 2001:105).

This means an ASI might see the original code humans made it with, in the same way we see colour or hear sound. Assuming that some of its original code was an attempt at solving the control problem, then this would amount to not much more than a suggestion to the AI that it should follow that code, in the same way that its other sensory information would be presenting it with possibilities for other actions it could take. Crucially, once a seed AI has evolved to become an ASI, we should not expect that the AI will treat any safety code we have implemented as privileged code (Yudkowsky, 2001:105), that is, code which the ASI cannot modify. Since, by definition, a seed AI is something which can modify any part of itself. This means premise (5) wouldn't be defeated by implementing some form of safety code that, for instance, destroys an AI as soon as it thinks about harming humans, since this very code could be modified by a self-modifying AI. I think it follows, that to implement code which *couldn't* be modified by a superintelligent self-modifying AI, would require that you are at least as intelligent as it, since it would amount to outsmarting the mind of an ASI. So, while we can't expect the code itself to do any of the work in limiting what the ASI can do, it is logically possible that the ASI take code meant to limit it as a *suggestion*, and then simply *consent* to following that suggestion. If an ASI does heed this early suggestion we make through code, it seems we have another counterexample to premise (5) such that something less intelligent has managed to limit what an ASI does and does not do. In a seed AI, we might program the AI to "never try to conceal [its] actions or cognitive state from [its] human programmers" (Yudkowsky, 2001:108). If the ASI that evolves from the seed AI does consent to following this instruction, then something less intelligent has limited what an ASI can do. Though any true ASI could just as easily ignore the instruction, like we would ignore a mosquito buzzing near our ear.

AIs that aren't seed AIs can also give consent. Consider a human brain which acquires superintelligence from a super-fast computer emulating it. This is an ASI which we will be unable to program any safety code for, since the mind we emulate will have exactly the same *programming* we do, though it would still be superintelligent. There is nothing that logically stops us from asking the superintelligent emulated mind: 'please don't harm humans.'

And likewise, there is nothing that logically stops the emulated mind from obeying that instruction. Perhaps it is unlikely for it to blindly follow the suggestion, just as we would be unlikely to obey a suggestion from an ant. Yet, something can be unlikely and still come true, and just one counterexample is all that's needed to break the identity claim in premise (5), that *all* things which control an ASI must be at least as intelligent as it. An ASI consenting to being limited does seem to be exactly that counterexample.

3. Significant & Insignificant Control

In section 1, I argued that a contradiction arises if we assume humans solve the control problem, and so there isn't a possible world where humans have or will ever successfully solve the control problem because that world would be an *impossible world* and "what is impossible is the case at no worlds" (Lewis, 2001:8). However, in section 2, I considered some strong objections in the way of thinking this. Both of these put pressure on premise (5) of my argument, which says that anything, human or otherwise, seeking to control a superintelligent agent, must be at least as intelligent as that agent, to successfully apply a given control method and limit the space of actions the agent can take. The first objection pointed out that it would be logically possible to limit what an ASI can do *before* it becomes an ASI. That is, implementing a control method in a seed AI which it carries through to future evolutions of itself. Similarly, the second objection showed that it is also logically possible for an ASI to consent, by luck or reason, to a control method which we want to implement. Perhaps the space of acceptable control methods an ASI would willingly consent to is small, but this makes it logically possible for us to pick something out of this space.

Before moving on, it's worth briefly looking at what a solution to the control problem would enable us to do. Most likely, we would use it to "subject AI [...] to positive moral and aesthetical values" (Sitnicki, 2018:§4). However, solving the control problem by itself would not be sufficient to achieve this task. Consider, as Yudkowsky and Bostrom (2014:331) do, what would happen if the ancient Greeks somehow had the ability to select the ethical values that would guide an ASI? Some of these values we might agree with but there are others that, today, we would find more than unattractive. For instance, thinking slavery is acceptable. This means that if the ancient Greeks did, in fact, create an ASI with their morals, then no matter how

much technological superiority their imaginary world experienced in relation to our actual world, they might never see or come to see slavery as being immoral. In part, this is due to the invulnerable nature of ASIs, like Dennet (2019:52) says, it's hard to imagine how we could dismantle an ASI without it switching to a mind at a backup location. So, there is a sense in which any ethical values we impose on an ASI will in turn be imposed on us, into the future. Crucially, the same moral *mistakes* we identify in the ancient Greeks would almost certainly be identified in any ASI that we built today, with our values. When our descendants look back to 2019, it's hard to imagine them not having the same reaction to some of our values, as we have to slavery (Graham, 2010).

More importantly, the lesson we are meant to draw from this, according to Yudkowsky and Bostrom (2014:331), is to not implement a control method in an ASI which only ensures that the ASI respects our current state-of-the-art morals, but rather we should implement a control method that allows an AI to evolve its morals, just as we evolve ours. In other words, a major and perhaps best use case for solving the control problem, will be to limit the space of possible actions an ASI can take not just once, but extending forward in time as long as the AI extends forward in time.

This distinction highlights an important fact about the control problem which shows why the objections against its impossibility fail. This is that, those objections only show that its logically possible for us to insignificantly control an ASI but do not show that its logically possible we significantly control an ASI.

Significant Control

It's useful to think about the control problem as the problem of how we can shrink the space of all possible actions an ASI can take. In an uncontrolled ASI, the space of all actions will include both stripping atoms on earth for resource acquisition and solving the remaining millennium prize problems. In a controlled ASI that space is smaller than in the uncontrolled ASI. This is consistent with the technical definition Bostrom (2014:127) gives, that the control problem is about how people who want to make an ASI, can ensure that *their* goals are met by the ASI. Looking at the control problem in these terms can easily lead us to the following premise:

> 3. A solution to the control problem will let the solver in some way limit what x [a truly superintelligent artificial intelligence] can do

But there is a sense in which this is incorrect. While it's true that solving the control problem will let you limit the space of available actions an ASI can take, the way that the space is limited is also important, because there are less and more significant ways the space of available actions can be limited. Suppose we create an ASI with the goal of collecting as many resources as possible, the ASI having a motivation to act on this goal, will limit the space of available actions it will take, so it does look as though, according to premise (3), if we successfully implemented that goal, we could be said to have solved the control problem, since we have limited what an ASI can do. Yet, there is also a sense in which that specific goal of resource acquisition limits the space of available actions insignificantly. Omohundro (2008) identifies resource acquisition as one of the basic drives *all* ASIs will have. So, limiting the space of available actions such that an ASI moves toward resource acquisition doesn't seem to be doing much, though under our current definitions of control, it is still a technical solution to the control problem.

Likewise, so long as we are correct about the speed of light, an ASI will not be able to causally affect anything outside our light cone, where a light cone of some object p is "all the points that can be connected to p by a straight ray of light." Since the speed of light has a limit, this is not infinitely far, which means that "anything outside the lightcone of p cannot causally interact with p" (Curiel, 2019). So just by building an ASI on Earth, we will be limiting the space of available actions it can take, in that, all of those actions will have to take place within our light cone. But again, though this means we would limit what the ASI can do, it does not seem like a significant limitation.

How then should we characterise what a *significant limitation* is? To rule out anything like the previous two examples, significant limitations cannot occur without our intervention. Meaning, if we were to build an ASI, any limitation that would occur even if we tried to implement *zero* limitations, is a limitation which is insignificant. Further, another feature that significant limitations should have is that they persist through time as long as the ASI persists through time. An ASI that is just limited for a day or an indeterminable amount of time, is not significant, since, while such limita-

tions would genuinely limit the space of available actions an ASI can take, it would only limit that space for a time less than the ASI is going to exist. Accordingly, any limitations we make could disappear at any moment. Though it might be possible to create a friendly AI based on such a solution to the control problem, it would be hard to convince anyone that this solution has neutralised the existential threat ASIs pose, if the solution were one that could fail before the ASI itself fails. With this idea of significant limitations in place we can reformulate premise (3) and (5) to:

> 3. A solution to the control problem will let the solver in some way **significantly** limit what x [a truly superintelligent artificial intelligence] can do
>
> 5. If y **significantly** limits what x can do, then y has a greater than or equal intelligence to x

These updated premises give us ground to reject the objection that an ASI could be controlled by humans, if we managed to give it some kind of goal that persists from when it is an AGI until it becomes an ASI. This is that, it's not clear a goal implemented in an AI before it becomes an ASI can persist indefinitely long. While Omohundro (2008:487) gave us reasons to think ASIs would act rationally, and Yudkowsky (2001:100) showed that an ASI wouldn't want to modify an initial goal set by us. Such that, given the goal of being nice to humans, because an AI wouldn't want to modify the goal, it will be present all the way until a seed AI evolves in to an ASI. However, what makes this counterexample fail, according to our new premises, is that it is really an example of *insignificant control* since all we can say for sure is that our limitation will be present in the ASI only as long as it wants it to be present. There is nothing that works to persist the goal in the AI besides the AI's own motivations. Yudkowsky (2001:222) even admits that "if the AI stops wanting to be Friendly, you've already lost." The limitation, then, is insignificant because we can't know that it will last as long as the ASI lasts, instead it is an open possibility how long the limitation will last. Whereas a significant limitation, will limit the space of possible actions an ASI can take as long as the ASI is functioning. So, to genuinely solve the control problem such that you can implement a lasting limitation does require that you are at least as intelligent as the ASI you are trying to limit.

One strong example that was meant to support this objection was the Gandhi example, where, we are meant to think that AIs would not want to change their goals because *even* humans would not want to. Incidentally, I think this example also fails. We can say that Gandhi happened to have a particularly strong moral motivation not to kill people, where a moral motivation is when a moral judgement motivates you to take some action (Rosati, 2016). In this case, Gandhi's moral judgement that killing people is wrong has *motivated* him to take the action of not killing people. We can't say that Gandhi's moral judgement *won't* change. So, if Gandhi morally judges that killing people is wrong, then we would expect him not to push a button that makes him judge killing people as right. However, nothing stops Gandhi from changing his initial moral judgement that killing people is wrong. He could, for instance, have his moral judgement unintentionally changed if his brain were damaged, and while this might not directly motivate him to kill people, his moral judgement could change enough for him to be motivated to push a button, that will make him judge killing people as not necessarily bad. Since, "when a person's judgment changes, [their] motivation tends to change [too]" (Rosati, 2016:§3.2). From this, I think we can conclude that, though Gandhi might be unlikely to push a button that will change his moral judgement, it's not *impossible* that he does push the button, if only because something unintentionally changes the way his moral judgements are produced.

Likewise, in the case of an ASI, we can't say that its goals will always remain stable because even human goals remain stable, human goals actually don't seem to. At least for humans, moral judgements can unintentionally change. In fact, it seems right to think that an ASI's goals could also change, if for nothing else than a cosmic ray causing a bit flip in its mind. All this means to say is that a goal implemented when an AGI is an AGI, cannot, just by it being the first goal it was programmed with, be guaranteed to persist through the entire life of its future ASI self.

For the same reasons, the case of an AI consenting to be limited also doesn't seem to be an example of a significant limitation. Even if an ASI were to consent to being limited in a way that isn't trivial, the ASI could withdraw this consent at any time, so we can't know that the limitation will last as long as the ASI does.

Conclusion

Talking about artificial superintelligences today is something several people will object to in and of itself. Dennett (2017:399) says that we won't have superintelligence within fifty years and that given this, we should not be worried about it now. Dennet might be right that it will take at least fifty years to reach superintelligence, if not more. According to Bostrom (2014:19) fifty percent of AI researchers think that general human artificial intelligence will only be reached by 2050, with ninety percent of AI researchers thinking it will arrive by 2090. I think, even if we assume there will be another very long AI winter, a period where little success is had with AI in general (Bostrom, 2014:7), then we would still be justified in worrying about the risks of ASIs now. Thinking that the risk is far away so it should be ignored in favour of other problems mistakes the nature of the risk that ASIs represent. That is, an existential risk, a risk which has the potential of killing all life on earth (Bostrom, 2002:§1.2). What is unique about existential risks is that "there is no opportunity to learn from errors" (Bostrom, 2002:§2). Further, because existential risks can cause such great damage, even with a low probability of occurring, they could still be put on par with the more tangible risks we face right now.

If we want to mitigate the existential risk of superintelligence, then solving the control problem is step zero. There are still many other issues, perhaps most we don't even know about, on the way to building a truly friendly ASI. So far, Bostrom (2019:1) has put our current general strategy toward new technologies best. That when it comes to new technologies, we have generally just hoped and got lucky that none of them have "invariably or by default [destroyed our] civilization." When it comes to ASIs, we cannot continue to use this strategy. If we do invent an AI which is capable of superintelligence, then it will not be enough to hope it is a technology which is beneficial rather than destructive. Instead, it is necessary that we make it a safe technology *before* it is invented. So far, AI researchers like Bostrom and Yudkowsky, have said making ASIs safe will involve solving the control problem. However, I have given reasons to think this is a logically impossible problem to solve for humans.

In section 1, I gave my argument for the control problem being logically impossible to solve, which said that, if we assume humans do solve the control problem this leads to a contradiction. But, in section 2, I looked at two strong objections to my argument which showed that it is possible hu-

mans could limit or control an ASI, if the limitation was implemented when the ASI were just generally intelligent and not superintelligent. Or if the ASI consented to being limited. However, in section 3, I showed that the types of limitations mentioned in section 2, would be insignificant because they could be ignored by the ASI at any time in the future, or they would be seemingly natural limitations that occur just by building an ASI. The kinds of control that the control problem is really concerned with is significant control, which is remains logically impossible for humans to implement.

References

Boden, M. A. (2016). Artificial Intelligence. In *Routledge Encyclopedia of Philosophy* (1st ed.). London: Routledge. https://www.rep.routledge.com/articles/thematic/artificial-intelligence/v-1.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology 9*(1). https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277–284. https://philpapers.org/archive/BOSEII.pdf.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies* (First edition ed.). Oxford: Oxford University Press.

Bostrom, N. (2019). The Vulnerable World Hypothesis. *Global Policy*, 1758–5899.12718. https://nickbostrom.com/papers/vulnerable.pdf.

Bringsjord, S. and N. S. Govindarajulu (2018). Artificial Intelligence. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/.

Curiel, E. (2019). Singularities and Black Holes. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/spacetime-singularities/lightcone.html.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Penguin UK.

Dennett, D. C. (2019). What Can We Do? In J. Brockman (Ed.), *Possible Minds: Twenty-Five Ways of Looking at AI* (1st ed.). New York: Penguin Press.

Dyson, G. (2019). The Third Law. In J. Brockman (Ed.), *Possible Minds: Twenty-Five Ways of Looking at AI* (1st ed.). New York: Penguin Press.

Fredericksen, M. A., Y. Zhang, M. L. Hazen, R. G. Loreto, C. A. Mangold, D. Z. Chen, and D. P. Hughes (2017). Three-dimensional visualization and a deep-learning model reveal complex fungal parasite networks in behaviorally manipulated ants. *Proceedings of the National Academy of Sciences 114*(47), 12590–12595. `http://www.pnas.org/lookup/doi/10.1073/pnas.1711673114`.

Graham, P. (2010). What You Can't Say. In *Hackers & Painters: Big Ideas from the Computer Age* (1st ed.). O'Reilly Media. `http://www.paulgraham.com/say.html`.

Harmon, K. (2012). Fungus that controls zombie-ants has own fungal stalker. *Nature News*. `http://www.nature.com/news/fungus-that-controls-zombie-ants-has-own-fungal-stalker-1.11787`.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books.

Lewis, D. K. (2001). *On the Plurality of Worlds*. Malden, Mass: Blackwell Publishers.

Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin (Eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, Number v. 171 in Frontiers in Artificial Intelligence and Applications, pp. 483–92. Amsterdam ; Washington, DC: IOS Press.

Robertson, T. and P. Atkins (2018). Essential vs. Accidental Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2018/entries/essential-accidental/`.

Rosati, C. S. (2016). Moral Motivation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/moral-motivation/.

Schroeder, M. (2016). Value Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2016/entries/value-theory/.

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv:1712.01815 [cs]*. http://arxiv.org/abs/1712.01815.

Sitnicki, I. (2018). Why AI shall emerge in the one of possible worlds? *AI & SOCIETY*. http://link.springer.com/10.1007/s00146-018-0833-9.

Yudkowsky, E. (2001). *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. California: The Machine Intelligence Research Institute. https://intelligence.org/files/CFAI.pdf.

Yudkowsky, E. (2002). The AI-Box Experiment. http://yudkowsky.net/singularity/aibox/.

Yudkowsky, E. (2007). Levels of Organization in General Intelligence. In B. Goertzel and C. Pennachin (Eds.), *Artificial General Intelligence*. Springer Science & Business Media. https://intelligence.org/files/LOGI.pdf.

Yudkowsky, E. and N. Bostrom (2014). The Ethics of Artificial Intelligence. In K. Frankish and W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence*. Cambridge, UK: Cambridge University Press. https://nickbostrom.com/ethics/artificial-intelligence.pdf.

Word Count: 10 901