Thinking in Advance About the Last Algorithm We Ever Need to Invent

Olle Häggström
Dept of Mathematical Sciences, Chalmers University of Technology, 412 96 Göteborg, Sweden, and Institute for Future Studies, Box 591, 101 31 Stockholm, Sweden
olleh@chalmers.se

Abstract

We survey current discussions about possibilities and risks associated with an artificial intelligence breakthrough on the level that puts humanity in the situation where we are no longer foremost on the planet in terms of general intelligence. The importance of thinking in advance about such an event is emphasized. Key issues include when and how suddenly superintelligence is likely to emerge, the goals and motivations of a superintelligent machine, and what we can do to improve the chances of a favorable outcome.

2012 ACM Subject Classification Computing methodologies → Philosophical/theoretical foundations of artificial intelligence

Keywords and phrases intelligence explosion, Omohundro–Bostrom theory, superintelligence

Digital Object Identifier 10.4230/LIPIcs.AofA.2018.5

Category Keynote Speakers

1 Introduction

In 1951, Alan Turing, in his *Intelligent machinery, a heretical theory* [41], anticipated many of the key ideas in current artificial intelligence (AI) futurology:

> My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely. [...] Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. [...] It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control.

One of Turing's collaborators at Beltchley Park, mathematician I.J. Good, later made a related prediction, in a famous passage [13] from which the title of the present paper is partly borrowed:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

The presently favored term for what Good called ultraintelligence is **superintelligence**: a superintelligent machine is one that by far exceeds human performance across the full range of relevant cognitive skills, including the mysterious-seeming quality we label creativity or the ability to think outside the box. Defining an agent's intelligence is of course not straightforward, and no strict definition will be given here, but it can be thought of informally as the ability to direct the world towards whatever goals the agent has. If a machine has at least human-level such ability across more or less the full range of domains encountered by humans, we speak of **artificial general intelligence (AGI)**, and if its general intelligence vastly exceeds that of humans, then it has superintelligence.

Is it really reasonable to expect superintelligence any time soon – let's say before the end of the present century? This is a highly controversial issue where expert opinions vary wildly, and while I accept that the question is wide open, I also hold – as the first of my two main claims in this paper – that the emergence of superintelligence is a sufficiently plausible scenario to warrant taking seriously. This claim is defended in Section 2 on the possibility in principle of superintelligence, and in Sections 3 and 4 on timelines.

The second main claim in this paper is that it is of great practical importance to think in advance about safety aspects of a superintelligence breakthrough, because if those aspects are ignored or otherwise mismanaged, the event might have catastrophic consequences to humanity. Such risks are discussed in Section 5, aided mainly by the Omohundro–Bostrom theory for instrumental vs final AI goals, which is explained in some detail. Ideas on how to ensure a more benign outcome are briefly discussed in Section 6, and Section 7 offers some concluding remarks.

2 The possibility in principle

Is a superintelligent machine possible in principle in the universe we inhabit? If a supernatural human soul – or something else in that vein – exists, then all bets are out the window, so I will ignore that possibility and instead focus on the case which is more amenable to rational argument: a physical world in which all high-level phenomena, including the human mind, are the result of particular arrangements of matter. Assuming this, the example of the human brain demonstrates that there are arrangements of matter that gives rise to human-level intelligence.

There are several independent ways to argue that the human brain is unlikely to be anywhere near an optimal arrangement of mater for producing intelligence. One is to point to the fact that our brain is the product of biological evolution, which viewed as an optimization algorithm is a rather primitive local search approach, which in a setting as complex as optimizing for intelligence is unlikely to find anything like a global optimum. Another thing to point at is the extreme slowness of the nervous system compared to how the same information processing might be carried out on a modern electronic computer. A third one is the many obvious miscalibrations and biases our brain has [12], that might be corrected for. See also Sotala [38] for further concrete examples of ways in which there is room for improvement upon human intelligence.

So there are good reasons to believe that there are physical arrangements of matter that produce intelligence far superior to the human brain, i.e., superintelligence. The argument so far does not show that it can be implemented on a digital computer, but if we accept the Church–Turing–Deutsch principle that a Turing-complete computing device can be used to simulate any physical process [9], then there is an algorithm out there that achieves superintelligence.

This argument is not entirely watertight, because if the algorithm is based on simulating the physical process on a very low level (say, the movement of elementary particles), then an implementation of it on a digital computer may turn out to be so slow that it cannot be recognized as superintelligent. But it seems plausible that more efficient implementations of the system's essential information processing should be possible. We note in passing that the level of detail with which a human brain needs to be implemented on a digital computer to capture its intelligence remains a highly open question [34].

While some uncertainty remains, considerations such as these strongly suggest the existence of algorithms that can be implemented on a digital computer to achieve superintelligence. Husfeldt [24] accepts the existence of such an algorithm, calls it *the monster in the library of Turing*, and suggests that it is prohibitively difficult to find such a monster. So even if we accept its existence, we should still be open to the possiblity that the answer to the question that the next section addresses – that of when we can expect a superintelligent machine – is "never". It might be that finding it requires – short of a thremodynamics-level mircle – astronomical (or larger) amounts of brute force search, so in the next sections's discussion on when to expect the emergence of superintelligence, time $t = \infty$ will be considered a genuine possibility.

3 When to expect superintelligence?

In view of the current surge of progress in AI for a wide range of applications such as speech synthesis [37], board games [36] and autonomous vehicles [25], it may be tempting to read this as a sign that AGI and superintelligence are just around the corner. We should not jump too quickly to such conclusions, however. Many commentators, including recently Jordan [26], emphasize a fundamental discontinuity between specialized AI applications and AGI – the former should not in general be understood as stepping stones towards the latter – and they may well be right. (On the other hand, see Yudkowsky [44] who points out that we do not have strong evidence to conclude that AGI and superintelligence are *not* around the corner.)

When looking at the history of AI, the contrast between the the extraordinary achievements in specialized AI applications and the much less impressive progress towards AGI is striking. It is sometimes claimed that the latter has been literally zero, but that seems to me a bit harsh. For instance, an AI was developed a few years ago that quickly learned to successfully play a range of Atari video games [29]. As I admitted in [19], this is of course a very far cry from the ability to handle the full range of tasks encountered by humans in the physical and social world we inhabit; nevertheless, it is a nonzero improvement upon having specialized skill in just a single video game. One possible path towards AGI, among many, might be a step-by-step expansion of the domain in which the machine is able to act intelligently.

We do not at present have very clear ideas on what approach to AI has the best potential for realizing AGI. The main driver behind the rapid progress we see today in various AI applications is the deep learning approach, which is essentially a rejuvenation and further development of old neural network techniques that used to yield unimpressive results but which in many cases work remarkably well today, thanks to faster machines and access to huge training data sets. It is not, however, written in stone that deep learning will retain its position as the dominant AI paradigm forever. Other potentially useful approaches that share the black box feature of deep learning include genetic programming mimicking biological evolution, and the brute force copying of the workings of the human brain in sufficient detail

to reproduce its behavior. This last possibility is advocated enthusiastically by Kurzweil [27] and discussed in more balanced fashion by Sandberg and Bostrom [34]. Alternatively, we might see a revival of the non-black box approach of GOFAI (Good Old-Fashioned AI) with explicit hand-coding of the machine's central concepts and reasoning procedures. Or perhaps some hitherto untried combination of these approaches, or something else entirely. It might be that none of these will ever yield AGI, but the reasonable stance seems to be to at least be open to the possibility that one of them might eventually accomplish that.

But when would that happen? This is highly uncertain, as illustrated by a survey by Müller and Bostrom [30] of estimates by the world's top 100 most cited AI researchers – eatimates that are spread out all over the present century, and beyond. Not only is the amount of between-individual differences large, the individually reported uncertainty ranges also tend to be broad. Among the 29 who responded, the median of their estimates for the time when human-level AGI can be expected to have arrived with probability 50% (given that "human scientific activity continues without major negative disruption") is 2050, with a median estimate of 50% for the probability that superintelligence emerges within 30 years later. More detailed but broadly consistent results are reported in the more recent survey by Grace et al. [14]. Yet another expert survey is reported in what looks like a deliberate attempt to downplay the importance of thinking ahead about AGI and superintelligence [11], but see [8] for an effective rebuttal.

The short answer to the question of when to expect superintelligence is that we do not know: experts are highly divided. In such a situation, it would be epistemically reckless to have a firm belief about if/when superintelligence will happen, rather than prudently and thoughtfully accepting that it may well happen within decades, or within centuries, or not at all.

Yet, it is quite common to hear, even among commentators for whom the label "AI expert" seems justified, dismissive attitudes towards the idea of a future superintelligence; Dubhashi and Lappin [10] and Bentley [3] are typical examples (see [20] for my fair and balanced response to the latter). Rarely or never do these commentators offer convincing arguments for their view. So one might wonder what the actual reasons for their view is, and although admittedly it is dubious to speculate on one's disputant's motives, I made a brave attempt in [17] to suggest an explanation for their stance in terms of what I decided to call vulgopopperianism, which I defined as the implicit attitude of someone who

> (a) is moderately familiar with Popperian theory of science, (b) is fond of the kind of asymmetry [appearing between the task of showing that all swans are white and showing that at least one non-white swan exists], and (c) rejoices in claiming, whenever he encounters two competing hypotheses one of which he for whatever reasons prefers, some asymmetry such that the entire (or almost the entire) burden of proof is on proving the other hypothesis, and insisting that until a conclusive such proof is presented, we can take for granted that the preferred hypothesis is correct.

The superintelligence timing case can for instance be concretized as a choice between two competing hypotheses (H1) and (H2), where (H1) is the hypothesis that achieving superintelligence is hard in the sense of not being attainable (other than possibly by extreme luck) by human technological progress by the year 2100. (H2) is the complementary hypothesis that achieving superintelligence is comparatively easy in the sense of being within reach of human technological progress (if allowed to continue unhampered) by 2100. A priori both hypotheses seem reasonably plausible, and the presently available evidence of one over the other is fairly weak (in both directions). This gives a vulgopopperian favoring (H1) the opportunity to focus on the shortage of evidence for (H2) and thus declare (H1) the winner –

while neglecting the shortage of evidence for (H1). This may be backed up with an analogy to the swan example: just like we stick to the "all swans are white" hypothesis until a non-white swan is encountered, we can stick with (H1) for as long as no superintelligence has been produced [17]. I believe this example would (or at least should) have made Popper nervous, because the idea behind his theory of falsificationism is to make science self-correcting [33], while in the case of stubbornly sticking to (H1) the desired self-correction (in case (H1) is wrong) is likely to materialize only the moment that superintelligence shows up and and it is too late for us to avert an AI apocalypse – a scenario whose plausibility I will argue for in Section 5.

4 How suddenly?

Related to, but distinct from, the question of *when* superintelligence can be expected, is that of *how sudden* its emergence from modest intelligence levels is likely to be. Bostrom [6] distinguishes between **slow takeoff** and **fast takeoff**, where the former happens over long time scales such as decades or centuries, and the latter over short time scales such as minutes, hours or days (he also speaks of the intermediate case of **moderate takeoff**, but for the present discussion it will suffice to contrast the two extreme cases). Fast takeoff is more or less synonymous with **the Singularity** (popularized in Kurzweil's 2005 book [27]) and **intelligence explosion** (the term coined by I.J. Good as quoted in Section 1, and the one that today is preferred by most AI futurologists). The practical importance of deciding whether slow or fast takeoff is the more likely scenario is mainly that the latter gives us less opportunity to adapt during the transition, making it even more important to prepare in advance for the event.

The idea that is most often held forth in favor of a fast takeoff is the **recursive self-improvement** suggested in the Good quote in Section 1. Once we have managed to create an AI that outperforms us in terms of general intelligence, we have in particular that this AI is better equipped than us to construct the next and improved generation of AI, which will in turn be even better at constructing the next AI after that, and so on in a rapidly accelerating spiral towards superintelligence. But is it obvious that this spiral will be rapidly accelerating? No, because alternatively the machine might quickly encounter some point of diminishing return – an "all the low-hanging fruit have already been picked" phenomenon. So the problem of deciding between fast and slow takeoff seems to remain open even if we can establish that a recursive self-improvement dynamic is likely.

Just like with the timing issue discussed in Section 3, our epistemic situation regarding how suddenly superintelligence can be expected to emerge is steeped in uncertainty. Still, I think we are at present a bit better equipped to deal with the suddenness issue than with the timing issue, because unlike for timing we have what seems like a promising theoretical framework for dealing with suddenness. In his seminal 2013 paper [43], Yudkowsky borrows from economics the concept of returns on reinvestment, frames the AI's self-improvement as a kind of cognitive reinvestment, and phrases the slow vs fast takeoff problem in terms of whether returns on cognitive reinvestment are increasing or decreasing in the intelligence level. Roughly, increasing returns leads to an intelligence explosion, while decreasing returns leaves the AI struggling to reach any higher in the tree than the low branches with no fruits left on them. From that insight, a way forward is to estimate returns on cognitive reinvestment based on various data sets, e.g, from the evolutionary history of *homo sapiens*, and think carefully about to what extent the results obtained generalize to an AI takeoff. Yudkowsky does some of this in [43], and leans tentatively towards the view that an intelligence explosion

is likely. This may be contrasted against the figures from the Müller–Bostrom survey [30] quoted in Section 3, which suggest that a majority of AI experts lean more towards a slow takeoff. I doubt, however, that most of these experts have thought as systematically and as hard about the issue as Yudkowsky.

5 Goals of the superintelligent AI: Omohundro–Bostrom theory

Consequences of an AGI breakthrough may turn out extremely beneficial to humanity, or they may turn out catastrophic. A favorite example of the latter – cartoonish on purpose to emphaisze that it is merely an example – is the so-called **Paperclip Armageddon**, which dates back at least to 2003 [4]. Imagine a paperclip factory, which is run by an advanced (but not yet superintelligent) AI, programmed to maximize paperclip production. Its computer engineers are continuously trying to improve it, and one day, more or less by accident, they manage to push the machine over the threshold where it enters the spiral of self-improvement causing an intelligence explosion. Coming out of the explosion is the world's first and only superintelligent AI. Having retained its goal of maximizing paperclip production, it promptly goes on to turn our entire planet (including us) into a giant heap of paperclips, followed by an expansion into outer space in order to turn the rest of the observable universe into paperclips. (For readers who feel repelled by the crude and seemingly farfetched character of Paperclip Armageddon, I recommend the more subtle and elaborate but no less frightening thought experiments offered by Armstrong [1] and Tegmark [40].)

Of course, AI futurology is not about randomly dreaming up weird scenarios, but about reasoning as rigorously as the topic admits about what is plausible and what is likely. The difficulty in evaluating whether an apocalypse along the lines of Paperclip Armageddon might really happen lies not so much in what a superintelligent machine would be *capable* of doing, but rather what it would be *motivated* to do. (For some vivid scenarios illustrating the capability of a superintelligent AI, see, e.g., [42], [6] and [40].) Currently the only game in town for going beyond mere speculations regarding a superintelligent AI's goals and motivations is what in my 2016 book [16] I decided to call **the Omohundro–Bostrom theory of final vs instrumental AI goals**, honoring key contributions by Omohundro [31, 32] and Bostrom [5, 6]. An agent's final goal is what the agent values as an end in itself rather than as a means towards achieving something else. An instrumental goal, in contrast, is one that is set up as a stepping stone towards another goal.

(Some philosophers, such as Searle [35], are fond of saying that this whole approach is confused, because computers cannot have goals. But the confusion is on their side, as even heat-seeking missiles and thermostats have goals in the relevant sense. See [15] for my detailed response to Searle.)

The two cornerstones of Omohundro–Bostrom theory are **the orthogonality thesis** and the **the instrumental convergence thesis**. We begin with the former.

> **The Orthogonality Thesis:** More or less any final goal is compatible with more or less arbitrarily high levels of intelligence.

In his original formulation, Bostrom [5] omits the qualifier "arbitrarily high" (writing instead "any"), but I prefer its inclusion so as not to have to bother with possible counterexamples that combine low intelligence with conceptually advanced goals. He does, however, include the qualifiers "more or less" (in both places), underlining the statement's lack of mathematical precision; it really does seem to be needed due to the kinds of counterexamples discussed towards the end of this section.

In response to the question "What will a superintelligent machine be inclined to do?", the Orthogonality Thesis on its own obviously isn't of much help in narrowing down from the useless answer "anything might happen". It does, however, serve as an antidote to naive (but fairly common; [22] is a typical example) anthropomorphisms such as "Paperclip Armageddon is impossible, since having such a stupid goal would directly contradict the very notion of superintelligence; surely someone who is superintelligent would realize that things like human welfare and ecosystem preservation are more important than monomanically producing ever-increasing numbers of paperclips," which conflate intelligence with goals. The Orthogonality Thesis helps remind us to distinguish between intelligence and goals.

More useful in terms of narrowing down on what a superintelligent machine can be expected to do is the Instrumental Convergence Thesis, in combination with a collection of concrete goals to which it applies.

> **The Instrumental Convergence Thesis:** There are several instrumental goals that are likely to be adopted by a sufficiently intelligent agent in order to pursue its final goal, for a wide range of final goals and a wide range of circumstances.

Omohundro [31] and Bostrom [5] list several instrumental goals that they argue to be in the range of applicability of the instrumental convergence thesis:

- **Self-preservation:** if you continue to exist and are up and running, you will be in a better position to work for your final goal compared to if you are turned off, so don't let anyone pull the plug on you!
- **Self-improvement:** improvements to one's own software and hardware design.
- **Acquisition of resources** such as hardware, but also things like money in case the agent operates in a world that is still dominated by the kind of economy we have today.
- **Goal integrity:** make sure your final goal remains intact.

The instrumental goal of self-improvement plays a special role in the theory of intelligence explosion discussed in Section 5, because it explains why, among the millions of other things it might decide to do, we should not be surprised to see the AI choose to work its way up the spiral of recursive self-improvement.

The value, for the purpose of pursuing a generic final goal, of the first three instrumental goals on the list is more or less self-explanatory, but the fourth item on the list – goal intrgrity – may warrant an explanation. As a simple example, imagine an AI with the goal of maximizing paperclip production, and suppose that, perhaps triggered by some external impulse, it starts to contemplate whether in fact ecosystem preservation might in fact be a preferable goal to pursue, compared to maximizing paperclip production. Should it stick to the old goal, or should it switch? In order to decide, it needs some criterion for which goal is the better one. Since it hasn't yet switched to the new goal, but is merely considering whether to do so, it still has the paperclip maximization goal, so the criterion will be: which goal is likely to lead to the larger number of paperclips? In all but some very contrived circumstances, paperclip maximimzation will win this comparison, so the AI will stick to that.

Equipped with Omohundro–Bostrom theory, we are in a position to understand that a scenario like Paperclip Armageddon is not as far-fetched as it first might seem. The Orthogonality Thesis helps us see that while paperclip maximization may seem bizarre *to us* (because we have other goals), it need not look that way *to the machine*, who may instead find goals like ecosystem preservation and promotion of human well-being utterly pointless. The instrumental goal of self-improvement helps explain why the paperclip maximizer might go through an intelligence explosion, and the instrumental goal of goal integrity explains why

the machine can be expected to come out of the intelligence explosion with its monomaniacal wish to produce paperclips intact.

A common objection to Paperclip Armageddon-like scenarios is that a superintelligent machine will understand that its original human programmers did not intend it to turn the observable universe into paperclips, and will therefore refrain from doing so. The mistake here is to take for granted that "do things that please your programmers" is among the machine's goals. Every programmer today knows that whenever there is a discrepancy between what the programmer intends and what appears literally in the computer code, it is the latter that counts. Omohundro–Bostrom theory predicts that principle to remain true for superintelligent machines. If that sounds like bad news, then perhaps a remedy might be to make "do things that please your programmers" the machine's final goal. Ideas in that spirit are in fact being considered in contemporary work on AI risk. More on that in the next section.

Before that, let me emphasize that while Omohundro–Bostrom theory is, for the time being, an indispensable tool for reasoning about consequences of an AGI breakthrough, it is also to some extent tentative. Its two cornerstones deal with messy concepts with fuzzy boundaries, and they do not (as yet, in their present form) deserve the same epistemic status as mathematical theorems that have been established once and for all. Therefore, predictions derived from the theory should be treated with some degree of epistemic humility (which is not to say that they can be dismissed out of hand). In my recent paper [18], I discuss a variety of challenges to the validity and range of applicability of Omohundro–Bostrom theory – in particular, the following three.

First, **self-referentiality**. Bostrom [5] points out that a superintelligent machine with the final goal of being stupid (properly specified) is unlikely to remain superintelligent for very long. Thus, for all practical purposes, the final goal of being stupid serves as a counterexample to the Orthogonality Thesis. Given one counterexample, how can we stop a wildfire of others? Some extra condition on the final goal needs to be found that excludes the stupidity example and whose inclusion makes the Orthogonality Thesis true. An obvious candidate is that the final goal cannot refer back to the machine itself, but the discussion in [18] points towards the task of defining such self-referentiality being highly problematic.

Second, **Tegmark's physics challenge**. Could other properties of a final goal, beyond self-referentiality, have the potential to invalidate the conclusion of the Orthogonality Thesis? A perhaps-too-obvious candidate is incoherence. What would it even *mean* for the machine to act towards an incoherent goal? Tegmark [39] suggests that the class of incoherent goals might be much bigger than we currently think:

> Suppose we program a friendly AI to maximize the number of humans whose souls go to heaven in the afterlife. First it tries things like increasing people's compassion and church attendance. But suppose it then attains a complete scientific understanding of humans and human consciousness, and discovers that there is no such thing as a soul. Now what? In the same way, it is possible that any other goal we give it based on our current understanding of the world (*"maximize the meaningfulness of human life"*, say) may eventually be discovered by the AI to be undefined.

Third, **human values are a mess**. If we believe that the Omohundro–Bostrom framework captures something important about the goal structure of a sufficiently intelligent agent, then we should also expect its neat dichotomy of final vs instrumental goals to be observable in such agents. The most intelligent agent we know of is *homo sapiens*, but the goals of a typical human do not seem to admit such a clearcut dichotomy [18].

6 AI Alignment

Various attempts have been made to avoid Turing's [41] conclusion (quoted in Section 1) that in the presence of superintelligent machines, "we should have to expect the machines to take control", but none of them seem to provide a clearcut solution. Probably the most studied such attempt is the so-called AI-in-a-box approach, which is to keep the machine boxed in and unable to influence the world other than via a narrow and carefully controlled communications channel. While this deserves further study, the present state-of-the-art seems to point in the direction that such boxing-in is extremely difficult and can be expected to work for at most a temporary and rather brief time period; see, e.g., [2] and [21].

It therefore makes sense to look into whether it is possible to accept that the superintelligent AI takes control and still get a favorable outcome (whatever that means). For that to happen, we need that the AI has goals that work out in our favor. Due to the instrumental goal of goal integrity, discussed in Section 5, it is unlikely that a superintelligent AI would allow us to tamper with its final goal, so the favorable goal needs to be installed into the AI *before* it attains superintelligence. This is the aim of the **AI Alignment** research program, formulated (under the alternative heading **Friendly AI**, which however is perhaps best avoided as it has an unnecessarily anthropomorphic ring to it) in Yudkowsky's seminal 2008 paper [42], and much discussed ever since; see, e.g., [6], [16] and [40].

Following Bostrom [6], we can think of AI Alignment as two problems: First, the difficult technical problem of how to encode whatever the desired goals are and install them into the AI – Bostrom calls this the **value loading problem** and "a research challenge worthy of some of the next generation's best mathematical talent". Second, the ethical problem of what the desired goals are, who gets to determine them, and via what procedure (democratic or otherwise). We probably do not want to leave it to a small group of AI developers in Silicon Valley or elsewhere to decide on the fate of humanity for the rest of eternity. Most thinkers in this field (including Yudkowsky [42] and Bostrom [6]) seem to agree that rather than explicitly hand-coding the values we wish the AI to have, an indirect approach is better, where somehow the AI is instructed to figure out what we want – or even better, what we would have wanted if we were more knowledgable and ethically mature, and had more time to think about it.

A key insight going back at least to Yudkowsky [42] is that human values are highly fragile, in the sense that getting them just a little bit wrong can bring catastrophic consequenecs in the mighty hands of a superintelligent AI. There may also be a tension between what is good for humanity and what is good in a less anthropocentric and possibly more objective sense: for instance, the goal "maximize the amount of hedonic utility in the world" might in a sense be very good for the universe, but is also likely to lead to the prompt extinction of humanity, as our bodies and brains are probably very far from optimizing the amount of hedonic utility per kilogram of matter.

Solving the AI Alignment problem should in my opinion be a high on the list of today's most urgent research tasks, but not for the reason that AGI and superintelligence would be likely to emerge during the next few years (although see [44]). Rather, even if they are decades away, the problem may well be so difficult that we need those decades to solve it, with little or no room for procrastination.

7 Concluding remarks

Let me conclude with the following remarks.

1. The reader may have noticed the discrepancy between Turing's [41] use of plural in talking about "machines [taking] control", and my use of singular when talking about the superintelligent AI. My choice of singular is due to what Bostrom [6] speaks of as "decisive strategic advantage": especially in case of a fast takeoff, the first machine to attain superintelligence can be expected to take control in such a way as to prevent other machines from challenging its power monopoly. But this outcome is not certain, and Bostrom devotes a chapter also to what he calls multipolar outcomes, with no such monopoly. Such an outcome might arise if AGI is first attained via brain emulations, at a time when our understanding of the human brain is still not good enough to enable us to tweak with the emulations much beyond what we already do to our brains today; Hanson [23] offers a rich and fascinating account of the many societal exotica that such a breakthrough might lead to.
2. Creating superintelligence is of course difficult, but creating superintelligence *and* AI Alignment may be even more difficult. This means that if several actors (companies or countries) compete over being the first (and probably only) one to create superintelligence, there may be an incentive to cut corners on the AI Alignment task or maybe even ignore it altogether. Such a situation would be terribly dangerous (see, e.g., Miller [28] and Cave and ÓhÉigeartaigh[7]), and should be avoided, e.g., by creating a spirit of international cooperation rather than competition. That is possibly easier said than done.
3. Apart from superintelligence there are many other problems about the future of AI that we urgently need to deal with, concering, e.g., integrity and mass surveliance, the social consequences of sexbot technology, autonomous weapons arms races, or the effects of automation on unemployment. It is sometimes suggested that the superintelligence discourse in AI futurology is a dangerous distraction from these other problems; see, e.g., Dubhashi and Lappin [10]. I agree that these other problems are extremely important, but I do not agree that this means that we should ignore superintelligence. It would be bad if we managed to navigate all those more down-to-earth societal problems with AI, only to end up being turned into into paperclips. We need to deal with all of these problems, including superintelligence.

References

1 Stuart Armstrong. *Smarter Than Us: The Rise of Machine Intelligence*. Machine Intelligence Research Institute, Berkeley, CA, 2014.

2 Stuart Armstrong, Anders Sandberg, and Nick Bostrom. Thinking inside the box: controlling and using an oracle AI. *Minds and Machines*, 22:299–324, 2012.

3 Peter Bentley. The three laws of artificial intelligence: Dispelling common myths. In *Should we fear artificial intelligence?*, pages 6–12. The EU Parliament's STOA (Science and Technology Options Assessment) committee, Brussels, 2018.

4 Nick Bostrom. Ethical issues in advanced artificial intelligence. In I. Smit et al, editor, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in A rtificial Intelligence, Vol. 2*, pages 12–17. International Institute of Advanced Studies in Systems Research and Cybernetics, 2003.

5 Nick Bostrom. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22:71–85, 2012.

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

7 Stephen Cave and Seán ÓhÉigeartaigh. An AI race for strategic advantage: rhetoric and risks. preprint, 2018.

8 Allan Dafoe and Stuart Russell. Yes, we are worried about the existential risk of artificial intelligence. *MIT Technology Review*, November 2016.

9 David Deutsch. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proceedings of the Royal Society A*, 400:97–117, 1985.

10 Devdatt Dubhashi and Shalom Lappin. AI dangers: imagined and real. *Communications of the ACM*, 60:43–45, 2016.

11 Oren Etzioni. No, the experts don't think superintelligent AI is a threat to humanity. *MIT Technology Review*, September 2016.

12 Thomas Gilovich, Dale Griffin, and Daniel Kahnemann. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge, UK, 2002.

13 I.J. Good. Speculations concerning the first ultraintelligent machine. In F. Alt and M. Rubinoff, editors, *Advances in Computers*, volume 6, 1965.

14 Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? Evidence from AI experts. arXiv:1705.08807, 2017.

15 Olle Häggström. Does the Chinese room argument preclude a robot uprising? OUPblog, Oxford University Press, January 2016.

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

17 Olle Häggström. Vulgopopperianism. Häggström hävdar, February 2017.

18 Olle Häggström. Challenges to the Omohundro–Bostrom framework for AI motivations. preprint, 2018.

19 Olle Häggström. Remarks on artificial intelligence and rational optimism. In *Should we fear artificial intelligence?*, pages 19–26. The EU Parliament's STOA (Science and Technology Options Assessment) committee, Brussels, 2018.

20 Olle Häggström. A spectacularly uneven AI report. Häggström hävdar, March 2018.

21 Olle Häggström. Strategies for an unfriendly oracle AI with reset button. In Roman Yampolskiy, editor, *Artificial Intelligence Safety and Security*. CRC Press, Boca Raton, FL, to appear, 2018.

22 Brett Hall. Superintelligence. Part 4: Irrational rationality. `http://www.bretthall.org/superintelligence-4.html`, 2016.

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

24 Thore Husfeldt. The monster in the library of Turing. `https://thorehusfeldt.files.wordpress.com/2015/04/bostrom.pdf`, 2015.

25 Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: problems, datasets and state-of-the-art. arXiv:1704.05519, 2017.

26 Michael Jordan. Artificial intelligence – the revolution that hasn't happened yet. *Medium*, April 2018.

27 Ray Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. Viking, New York, 2005.

28 James Miller. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*. Benbella, Dallas, TX, 2012.

29 Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

30 Vincent Müller and Nick Bostrom. Future progress in artificial intelligence: a survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*, pages 553–571. Springer, New York, 2016.

31 Stephen Omohundro. The basic AI drives. In P. Wang, B. Goertzel, and S. Franklin, editors, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pages 483–492. IOS, Amsterdam, 2008.

32 Stephen Omohundro. Rational artificial intelligence for the greater good. In A. Eden, J. Moor, J. Søraker, and E. Stenhart, editors, *Singularity Hypotheses: A Scientific and Philosophical Assessment*, pages 161–175. Springer, New York, 2012.

33 Karl Popper. *The Logic of Scientific Discovery.* Hutchinson, London, 1959.

34 Anders Sandberg and Nick Bostrom. Whole brain emulation: a roadmap. Future of Humanity Institute technical report #2008-3, 2008.

35 John Searle. What your computer can't know. *New York Review of Books*, October 2014.

36 David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv:1712.01815, 2017.

37 R.J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive apeech aynthesis with Tacotron. arXiv:1803.09047, 2018.

38 Kaj Sotala. How feasible is the rapid development of artificial superintelligence? *Physica Scripta*, 92:113001, 2017.

39 Max Tegmark. Friendly artificial intelligence: the physics challenge. arXiv:1409.0813, 2014.

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

41 Alan Turing. Intelligent machinery: a heretical theory. BBC, 1951.

42 Eliezer Yudkowsky. Artificial intelligence as a positive and negative factor in global risk. In Nick Bostrom and Milan Cirkovic, editors, *Global Catastrophic Risks*, pages 308–345. Oxford University Press, Oxford, 2008.

43 Eliezer Yudkowsky. Intelligence explosion miroeconomics. Machine Intelligence Research Institute, 2013.

44 Eliezer Yudkowsky. There's no fire alarm for artificial general intelligence. Machine Intelligence Research Institute, 2017.