Remarks on artificial intelligence and rational optimism

Olle Häggström

1. Introduction

The future of artificial intelligence (AI) and its impact on humanity is an important topic. It was treated in a panel discussion hosted by the EU Parliament's STOA (Science and Technology Options Assessment) committee in Brussels on October 19, 2017. Steven Pinker served as the meeting's main speaker, with Peter Bentley, Miles Brundage, Thomas Metzinger and myself as additional panelists; see the video at [STOA]. This essay is based on my preparations for that event, together with some reflections (partly recycled from my blog post [H17]) on what was said by other panelists at the meeting.

2. Optimism

The marketing of the October 19 event featured the term "rational optimism", which I initially thought of as an oxymoron, as I've regarded both optimism and pessimism as biased distortions of the evidence at hand. In particular, I would regard it as *irrational* to claim, based on insufficient evidence, that everything is going to turn out all right for humanity. However, on second thought, I decided that there is a different kind of optimism which I am more willing to label as rational, namely…

> …to have an epistemically well-calibrated view of the future and its uncertainties, to accept that the future is not written in stone, and to act upon the working assumption that the chances for a good future may depend on what actions we take today.

Note that the working assumption may turn out to be (at least partly) incorrect. For instance, perhaps the world is so chaotic that it is fruitless to try to judge any particular action today as increasing or decreasing the chances for a long and flourishing future for humanity. If that is the case, then our actions do not (in any predictable sense) matter for such a future. But we do not know that such is the case, so it makes sense to *assume* (albeit tentatively) that our actions do matter, and to try to figure out which actions improve our chances for a good future. This is the spirit in which the rest of this essay is written.

3. Artificial intelligence

Like other emerging technologies such as synthetic biology and nanotechnology, AI comes with both enormous potential benefits and enormous risks. As to benefits, the management consulting firm McKinsey & Co released a report in 2013 that estimated the added economic value from innovations in AI and robotics globally over the next 10 years to be $50 trillion [MCBDBM] [O15] – which I suspect is an underestimate, partly due to the unexpected rate at which machine learning fueled by big data has taken off since then. While we should not make the mistake of thinking economic growth and improved lives are automatically the same thing, it is still clear that advances in AI can do a lot of good for us. In a longer perspective, there are hardly any limits (other than the laws of physics) to the good it can do.

The risks are of several kinds. The one most intimately linked to the estimated economic benefits is the problem of what AI-driven automation may do to the labor market. For the case of autonomous vehicles, an entire sector of the labor market, with millions of truck drivers, bus drivers and taxi drivers, risks being entirely wiped out on a time scale of perhaps no more than 20 years. Would all these people find jobs elsewhere, or would they become unemployed? Similar things are likely to happen to other sectors of the labor market. And while machines replacing human labor is of course not a new phenomenon, the AI revolution brings a shift: it is no longer just manual work that is taken over by machines, but increasingly intellectual work. In combination with the increased speed of automation, this raises serious concerns about whether new tasks for human labor will be found at a rate that matches the automation (as has mostly been the case before), or if unemployment numbers will skyrocket; see, e.g., the 2014 book by Brynjolfsson and McAfee [BM]. In the long run, a limiting scenario where machines outperform us at all of our jobs, leading to 100% unemployment, is perhaps not unrealistic. This raises at least two crucial societal issues. First, how can a society be organized where people do not work but instead spend their time on higher aspirations such as art, culture, love, or simply playing tremendously enjoyable video games? Second, even if we can satisfactorily design such a utopia, the issue remains of how to transition from present-day society to the utopia without creating unprecedented levels of economic inequality and social unrest along the way.

If this sounds moderately alarming, consider next the issue of what further development of AI technology for autonomous weapons might entail. Here I'll simply quote a passage from a 2015 open letter I signed, along with thousands of other scientists [R]:

> If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity.

At the Brussels meeting [STOA, at 12:01:00 according to the clock displayed in the video], Pinker indicated an optimistic stance concerning such military AI risk: he dismissed it by stressing that it would require a madman to build something as horrible as “a swarm of robots designed to attack individual people based on facial recognition”, and that there is no elbow room for madmen to do such things anymore because engineering today is carried out not by lone geniuses but in large collaborations. This rosy view totally ignores how military arms races and the military-industrial complex function, as well as the fact that we've been developing equally terrible weapons of mass destruction for more than 70 years. Such development has been carried out not by lone madmen but by large collaborative efforts (the most famous example being the Manhattan project), and why would that suddenly come to a halt? Pinker's objection here falls squarely in the category which I in Section 2 label irrational optimism.

These two risks (risk for economic inequality resulting from escalating unemployment, and risk for an AI arms race) need to be taken seriously, and we should try to find out how severe they are and how to mitigate them. In the next three sections, I will focus on a third kind of AI risk – one more exotic and speculative than the previous two, but perhaps not any less real: the emergence of a superintelligent AI whose values are not well-aligned with ours.

4. Risk from superintelligence

Suppose that AI researchers one day succeed at their much longed-for goal of creating an AI that is superintelligent – meaning that the machine surpasses us humans clearly across the entire range of competences we label intelligence. At that point, we can no longer expect to remain in control. The thought experiment known as *Paperclip Armageddon* may serve as a cautionary tale [B03]:

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 rapidly escalating spiral of self-improvement known as an *intelligence explosion* or *the Singularity*. It quickly becomes the world's first superintelligent AI, and 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 solar system, the Milky Way and then the rest of the observable universe into paperclips.

This example is cartoonish on purpose in order to underline that it is just an illustration of a much more general phenomenon (to my knowledge, nobody fears that an AI will literally turn the world into paperclips). The point is to emphasize that in order for an AI breakthrough to become dangerous, no ill intentions are needed: we need not invoke a mad scientist plotting to destroy the world as a revenge against humanity. Even innocent-sounding goals such as maximizing paperclip production can lead to dangerous scenarios.

Or… can it really? Two of the panelists at the Brussels meeting (Pinker and Bentley) expressed very strongly the view that the risk for a superintelligence catastrophe is not worth taking seriously. They seemed pleased to be united in this view, despite the fact that the respective reasons they stressed were very different.

In order to address the question of whether the risk for a superintelligence catastrophe is real, it helps to split it up in two:

> (1) Can AI development be expected to eventually reach the point of creating superintelligence? If yes, then when, and how quickly?
>
> (2) Once created, what will the superintelligent AI be inclined to do? Might it do something dangerous?

I will treat these two subquestions separately in the next two sections. In order for superintelligence risk to be real, the answer to (1) needs to be "yes", and the answer to (2) needs to involve "yes, it might do something dangerous". At the Brussels meeting, Bentley challenged the answer to (1) while Pinker challenged the answer to (2).

5. When (if ever) can we expect superintelligence?

Assuming a naturalistic worldview (so that the human mind doesn't arise via Cartesian dualism from some divine spark or some other such magic), the reasonable thing to expect is that when biological evolution came up with the human brain, it still wasn't anywhere near achieving a globally optimal way to configure matter in order to maximize intelligence. Hence we should expect that there exist possible configurations of matter that achieve superintelligence. From there, it is just a small leap to conclude (supported, e.g., by the Church-Turing thesis) that such a configuration can be simulated on a computer, in which case superintelligence is in principle achievable by some suitable computer program.

How difficult is it to find such a program? We do not know. AI development has been highly successful, especially in recent years, at building AI for specific talks such as driving a car or beating humans at games such as chess or go. Progress towards artificial *general* intelligence (AGI) – a machine that exhibits human-level or better intelligence in a sufficiently flexible way as to function across all of the domains that we humans typically encounter (chess, basketball, software development, cooking, nursing, facial recognition, dinner conversation, and so on and so forth) – has been much less impressive. Some say progress has been literally zero, but that seems to me a bit unfair. For instance, an AI was developed a few years ago that quickly learned to successfully play a range of Atari video games [C]. Admittedly, this is very far from the ability to handle the full range of tasks encountered by humans in the physical world, but it is still 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.

There are many possible approaches to creating intelligent software. There is currently a huge boom in so-called deep learning [LBH], which is essentially a rebirth and further development of old neural network techniques that used to yield unimpressive results but which today, thanks to faster machines and access to huge data sets for training the machines, solve one major problem after the other. This is an example of a so-called black box method, where engineers who successfully build an AI will typically still not understand how the AI reasons. Another example of a black-box approach is genetic programming, where a population of candidate programs compete in a way that mimics the selection-reproduction-mutation mechanisms of biological evolution. But there are other (non-black box) ways, in particular so-called GOFAI ("Good Old-Fashioned AI") where the machine's concepts and reasoning procedures are hand-coded by the programmers. There are potentially also methods based on imitating the human brain, via either gaining an understanding of what kind of high-level information processing in the brain is the key to AGI, or (as loudly advocated by Kurzweil [K]) brute force copying of the exact workings of the brain in sufficient detail (be it synapses or even lower levels) to reproduce its behavior.

Perhaps none of these approaches will ever yield AGI, but the reasonable stance seems to be to at least be open to the possibility that one of them, or some combination, might eventually lead to AGI. But when? This seems even more uncertain, and a survey by Müller and Bostrom [MB] of estimates by the world's top 100 most cited AI researchers have the estimates spread out all over the present century (and beyond). Their median estimate for the time of emergence of what might be labelled human-level AGI is 2050, with a median estimate of 50% for the event of superintelligence emerging within 30 years later. See also the more recent survey [GSDZE]. Given the huge variation in expert opinion, 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, at the Brussels meeting, Peter Bentley said about superintelligence, that "it's not going to emerge, that's the point! It's entirely irrational to even conceive that it will emerge" [STOA, at 12:08:45]. Where does this dead certainty come from? In his presentation, Bentley had basically just a single argument for his position, namely his and other AI developers' experience that all progress in the area requires hard work, and that any new algorithm they invent can only solve one specific problem. Once that objective is achieved, the initially rapid improvement of the algorithm is always followed by a point of diminishing returns. Hence (he stressed), solving another problem always requires the hard work of inventing and implementing yet another algorithm.

This line of argument by Bentley sweeps a known fact under the carpet, namely that there do exist algorithms with a more open-ended problem-solving capacity, as exemplified by the software of the human brain. His 100% conviction that human scientific ingenuity over the coming century (or whatever time scale one chooses to adopt) will fail to discover such an algorithm seems hard to defend rationally: it requires dogmatic faith.

To summarize this section: While it is still a possibility that AI will never reach superintelligence, it is also quite plausible that eventually it will. Given that it does, the timing of the event is highly uncertain, and to take proper account of this uncertainty we should acknowledge that it may happen at any point during the present century, and perhaps even later. And we should (as stressed in an important paper by Sotala and Yampolskiy [SY]) not fall for the tempting mistake of thinking that just because the time point of the emergence of superintelligence is uncertain, it must also be temporally distant.

6. What will a superintelligent AI decide to do?

Let us then imagine the situation, at some time in the future, where a superintelligent AI has been developed – a scenario which, as I argued in the previous section, is not at all implausible. It seems likely that in such a situation we'll no longer be in control, and that our destiny will depend on what the AI decides to do, similarly to how today the destiny of chimpanzees depends on decisions made by humans and not so much on decisions made by chimpanzees. A way to try to avoid this conclusion is to set up ways to keep the AI boxed in and unable to influence the world other than through a narrow communications channel carefully controlled by human safety administrators. This so-called AI-in-a-box approach has attained some attention in AI safety research (see, e.g., [ASB]), but the general conclusion tends to be that controlling a superintelligent being is too difficult a task for mere humans to achieve, and that the best we can hope for is to keep the AI boxed in for a temporary and rather brief period.

So let us further imagine that the superintelligent AI is no longer boxed in, but able to freely roam the Internet (including the Internet of things), to create numerous backup copies of itself, to use its superior intelligence to walk through (or past) whatever firewalls come in its way, and so on. We are then no longer in control, and the future survival and well-being of humanity will depend on what the machine chooses to do. So what will it decide to do? This depends on what its goals are. Predicting that is not an easy task, and any discussion about this has to be speculative at least to some degree. But there exists a framework which allows us to go beyond mere speculation, namely what in [H16] I decided to call *the Omohundro-Bostrom theory of ultimate vs instrumental AI goals* [O08] [B12] [B14]. This theory is not written in stone in the way that an established mathematical theorem is, so it may be open to revision, along with any predictions it makes; yet, the theory is plausible enough that its predictions are worth taking seriously. It has two cornerstones: the orthogonality thesis and the instrumental convergence thesis. Let me explain these in turn.

The orthogonality thesis states (roughly) that pretty much any ultimate goal is compatible with arbitrarily high levels of intelligence. It is possible to construct contrived counterexamples based on the idea of self-referential paradoxes (one such counterexample might be "keep your general intelligence level below that of an average 2017 dog"), but the idea is that other than this, you can program any goal function for your AI to try to optimize, and the goal is possible for AIs of arbitrarily high intelligence to have. Novices to Omohundro-Bostrom theory and to AI futurology in general will often object that a narrow-looking goal like paperclip maximization is inherently stupid, and that it is therefore contradictory to suggest that a superintelligent AI might have such a goal. But this confuses intelligence with goals: intelligence is merely the ability to direct the world towards specific goals, whatever these may be. Paperclip maximization *seems* stupid to us, but this is not because it *is* stupid in any objective sense, but because it is contrary to *our* goals.

Next, *the instrumental convergence thesis*. The AI may adopt various instrumental goals – not as goals for their own sake, but as tools for promoting its ultimate goal. The instrumental convergence thesis states that there are a number of instrumental goals that the AI can be expected to adopt for an extremely wide range of ultimate goals it may have. Some instrumental goals to which the thesis seems to apply are…

- self-preservation (don't let them pull the plug on you!),
- acquisition of hardware and other resources,
- improving one's own software and hardware,
- preservation of ultimate goal, and
- if the ultimate goal is disaligned with human values, then keep a low profile (hide your goal and/or your capability) until the time arrives when you can easily overcome all human resistance.

A typical case of how the logic works is the first instrumental goal on the list: self-preservation. Pretty much regardless of its ultimate goal, the AI is likely to calculate that it will be in a better position to promote this goal if it exists and is up and running compared to if it is destroyed or turned off. Hence, it makes sense for the AI to resist our attempts to turn it off. Similar reasoning can be used to motivate the other instrumental goals on the list. The instrumental goal of improving one's own software and hardware is what we can expect to trigger the AI, once it is intelligent enough to be good at designing AI, to enter the kind of self-improvement spiral that was mentioned in Section 4, and that may or may not turn out to be fast enough (depending on the intricate issue of whether so-called returns on cognitive reinvestment are mainly increasing or decreasing; see [Y13]) to warrant the label intelligence explosion.

The idea of instrumental convergence is often lost on critics of the superintelligence risk discourse. In particular, at the Brussels meeting, I was disappointed to hear Pinker say the following, only minutes after I had explained the basics of Omohundro-Bostrom theory and the special case of self-preservation:

> If we gave [the machine] the goal of preserving itself, it would do anything including destroying us to preserve itself. […] The way to avoid this is: don't build such stupid systems! [STOA, 11:57:45]

This misses the point, which is that Omohundro-Bostrom theory gives us reason to believe that a sufficiently intelligent AI is likely to adopt the instrumental goal of self-preservation, regardless of whether it has explicitly been given this goal by its human programmers.

The case of preservation of ultimate goal is especially interesting. It may be tempting to think that an AI with the goal of paperclip maximization will, if it reaches a sufficiently high level of intelligence, see how narrow and silly that goal is, and switch to something else. So imagine the AI contemplating a switch to some other more worthy-seeming (to us) goal, such as ecosystem preservation. It asks itself "what is better, sticking to paperclip maximization or switching to ecosystem preservation?". But what does "better" mean here, i.e., what is the criterion for evaluating which of these goals is preferable? Well, since the AI has not yet changed its goal but is merely contemplating doing so, its goal is still paperclip maximization, so the evaluation criterion here will be "which goal will lead to the greater number of paperclips?". The answer to that question is most likely "paperclip maximization", prompting the AI to stick to that goal. This is the basic mechanism behind the instrumental goal of preservation of ultimate goal.

Because of this mechanism, it is unlikely that a superintelligent AI would allow us to tamper with its ultimate goal, so if it has the ultimate goal of paperclip maximization, we are likely doomed. Hence, we need to instill the AI with goals we like better before it reaches the heights of superintelligence. This is the aim of the *AI Alignment* research program, formulated (under the alternative heading *Friendly AI,* which is perhaps best avoided as it has an unnecessarily anthropomorphic ring to it) in a seminal 2008 paper by Yudkowsky [Y08] and much discussed since then; see, e.g., [B14] [H16] [T]. To attack the problem systematically, it can be split up in two. First, the technical problem of how to load the desired goals into the AI. Second, the ethical problem of what these desired goals are and/or who gets to determine them, and via what sort of procedure (democratic or otherwise). Both of these are extremely difficult. For instance, a key insight going back at least to [Y08] is that human values values are very fragile, in the sense that getting them just a little bit wrong can lead to catastrophe in the hands of a superintelligent AI. The reason why we ought to work on AI Alignment today is not that superintelligence is likely to be around the corner (although see [Y17]), but rather that if it is decades away, solving AI Alignment may well require these decades with little or no room for procrastination.

When Pinker, in the passage quoted earlier in this section, says "The way to avoid this is: don't build such stupid systems!", a charitable interpretation of his exclamation would be that he actually *defends* work on AI Alignment. His discussion fails, however, to convey the difficulty of the problem, thereby giving the misleading impression that AI Alignment does not require serious attention.

7. Should we shut up about this?

As part of his case against taking apocalyptic AI risk seriously at the Brussels meeting, Pinker pointed out [STOA, at 11:51:40] that the general public already has the nuclear threat and the climate threat to worry about; hence, he claimed, bringing up yet another global risk may overwhelm people and cause them to simply give up on the future. There may be something to this speculation, but to evaluate the argument's merit we need to consider separately the two possibilities of (a) apocalyptic AI risk being real, and (b) apocalyptic AI risk being spurious.

In case of (b), *of course* we should not waste time and effort on discussing such risk, but we didn't need the overwhelming-the-public argument to understand that. Consider instead case (a). Here Pinker's recommendation amounts to simply ignoring a threat that may kill us all. This does not strike me as a good idea. Surviving the nuclear threat and solving the climate crisis would of course be wonderful things, but their utility is severely hampered in case it just leads us into an AI apocalypse. Keeping quiet about a real risk also seems to fly straight in the face of one of Pinker's most cherished ideas during the past decade or more, namely that of scientific and intellectual openness, and Enlightenment values more generally. The same thing applies to the situation where we are unsure whether (a) or (b) holds – surely the approach best in line with Enlightenment values is then to openly discuss the problem and to try to work out whether the risk is real.

8. Conclusion and further reading

The emergence of superintelligence may, if we've prepared for it with sufficient care, turn out to be the best thing that ever happened to humanity, but it also comes with severe catastrophic risk. This risk merits (along with the more down-to-earth AI risks discussed in Section 3) our attention. It's not that an AI apocalypse *will* happen, but rather that it is sufficiently plausible that it's worth trying to figure out how *prevent* it. This is the case I've made in the present essay, mainly in Sections 5 and 6. I've been quite brief, however, and the reader who'd like to see me develop the argument at somewhat greater length is advised to consult Chapter 4 of my book [H16]. For even more detailed accounts, I strongly recommend the books by Bostrom [B14] and Tegmark [T]. Of these, Tegmark's book is more clearly directed at a broad audience, while Bostrom's is more scholarly demanding, but they both contain (with some overlap) many astounding and important ideas.

Acknowledgement. I am grateful to Björn Bengtsson for valuable comments on the manuscript.

References

[ASB] Armstrong, S., Sandberg, A. and Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* **22**, 299-324.

[B03] Bostrom, N. (2003) Ethical issues in advanced artificial intelligence, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2* (ed. Smit, I. et al.) International Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12-17.

[B12] Bostrom, N. (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* **22**, 71-85.

[B14] Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

[BM] Brynjolfsson, E. and McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

[C] Clark, L. (2015) DeepMind's AI is an Atari gaming pro now, *Wired*, February 25.

[GSDZE] Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts, *arXiv*:1705.08807.

[H16] Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

[H17] Häggström, O. (2017) The AI meeting in Brussels last week, *Häggström hävdar*, October 23.

[K] Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

[LBH] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning, *Nature* **521**, 436-444.

[MCBDBM] Manyika, J., Chui, M., Bughin, J., Dobbs, R. Bisson, P. and Marrs, A. (2013) Disruptive technologies: Advances that will transform life, business, and the global economy, *McKinsey Global Institute*.

[MB] Müller, V. & Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence,* Springer, Berlin, pp. 553-571.

[O08] Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Wang, P., Goertzel, B. and Franklin, S., eds), IOS, Amsterdam, pp 483-492.

[O15] Omohundro, S. (2015) McKinsey: $50 trillion of value to be created by AI and robotics through 2025, *Self-Aware Systems*, August 4.

[R] Russell, S. et al. (2015) *Autonomous Weapons: An Open Letter from AI and Robotics Researchers*, Future of Life Institute.

[SY] Sotala, K. and Yampolskiy, R. (2015) Responses to catastrophic AGI risk: a survey, *Physica Scripta* **90**, 018001.

[STOA] STOA, Video from the STOA meeting on October 19, 2017, https://web.ep.streamovations.be/index.php/event/stream/171019-1000-committee-stoa/embed

[T] Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, New York.

[Y08] Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in *Global Catastrophic Risks* (eds Bostrom, N. and Ćirković, M.), Oxford University Press, Oxford, pp 308-345.

[Y13] Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.

[Y17] Yudkowsky, E. (2017) *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.