Ban Advanced Future AI Before It Is Too Late?

The Existential Risk Observatory On Advanced AI

James Fleming[1]

Abstract

How scary is advanced AI? So scary it seems that Otto Barten quit his career as a physicist and sustainable energy engineer to set up a think tank called the Existential Risk Observatory dedicated to stopping it from ever emerging. In this article, part of a series in the lead up to this year's Ron McCallum Debate on AI and work, the Journal of Work and Ideas (JWI) picks Otto's brains on the dangers of the rapid and largely unchecked advances in AI, and what we should do about them. You might be wondering if the promise of sophisticated AI to solve poverty, address climate change and to enhance science and medicine is too great to warrant pausing its development. However, Otto does not think so – he believes that the dangers are simply too great. When should we pause progress in AI sophistication? In his view, possibly, already in a few years.

Otto is the founder and director of the Existential Risk Observatory, an Amsterdam-based non-profit aiming to reduce existential risk, especially from AI, by informing the public debate. He has a background in physics, sustainable energy technology, and data science.

Key words

Artificial intelligence, existential risk, AI safety, AI governance, AI alignment, technological unemployment, international regulation, AI moratorium, Superintelligence, work

[1] James Fleming is the Executive Director of AIER.

JWI: What first sparked your concerns about the existential risks posed by advanced AI systems?

Otto Barten: *I attended a lecture in London by a futurist called Anders Sandberg about five years ago, who was then working for the Future of Humanity Institute at the University of Oxford. Using exploratory engineering, he tried to say something meaningful about what could be our long-term future. One of his conclusions, with which I agree, was that it was not unlikely that humanity would become extinct because of technology we are building ourselves. Current examples of such man-made extinction risks include atomic weapons, arguably climate change (as an externality of steam and combustion engines), and the possibility of pandemics created by using biotechnology.*

In the relatively near future (this century), it seems likely that we will invent human-level AI as well, which also poses a major risk of human extinction or a permanent loss of control. I walked away from this lecture shocked that we may currently be on the way towards causing our own extinction, and that we are not structurally trying to avoid most of those scenarios. This is what prompted me to start the Existential Risk Observatory.

JWI: You've criticised AI alignment efforts as insufficient for preventing catastrophic outcomes. Could you elaborate on the key technical and philosophical challenges of aligning AI with human values?

Otto Barten: *Gladly. The concept of AI alignment has been proposed by researchers as a solution against us losing control over AI, which could otherwise, according to many experts, result in human extinction or a permanent loss of control over our future. Their thinking was that it would likely be impossible to control AI beyond a certain capabilities level. However, if we would align the superhuman AI, there would be no need for control since the uncontrollable, superhuman AI would always act in accordance with our values.*

I have always thought this was an extremely dangerous strategy. First, humanity does not have universally agreed-upon values. Second, attempts by philosophers to put human values into words are

sketchy at best. If we would have an AI that we would not be able to control anymore, and it would actually implement, globally, down to molecular level, some AI interpretation of the words that some philosophers once wrote down and described as 'our values', it would almost certainly lead to a dystopia worse than anything we have seen so far. There are somewhat less naive interpretations of alignment, but I have not yet heard of any convincing way to align AIs which we might lose control over.

Apart from this perhaps philosophical problem, there is the technical problem of how to reliably make a superhuman AI do what we ask it to do. This problem is generally seen as unsolved by alignment researchers and progress on it, though not absent, seems much slower than progress on AI's capabilities.

JWI: There is no global consensus on what values AI should adopt. How do you think this lack of agreement on human values complicates the development of superintelligent AI?

Otto Barten: *The development of superintelligent AI itself is unfortunately not slowed down by this fact, but the development of a feasible alignment strategy is. In my opinion, 'human values' is a concept not meaningful enough to be useful in aligning an all-powerful AI. We really have no idea what we mean by saying these words. Human individuals do have preferences for one state over another, and those preferences can to an extent be measured. An approach I would be somewhat less pessimistic about is to somehow aggregate individual human preferences and act upon the outcome. Democracy is one way to aggregate human preferences. A democratic outcome is indeed one option which could be used as input for an AI.*

Perhaps we can think of better preference aggregation options, possibly AI-powered. This is all assuming technical alignment will get solved, which looks unlikely at the moment. In a situation where technical alignment gets solved and individual preference aggregation is used, there could be a tension between doing what humans want and doing what may be good for us in the long run. Also, decisions

will need to be made about who to give voting rights. How should we include the interests of children, and those of non-human animals?

Realistically, we have only begun to scratch the surface of all the questions that need answering in a world with superhuman AI, and not answering them well can be existentially dangerous. The long-term trend in new technologies is that they get ever-more powerful: from fire to the wheel, to the steam engine, to nuclear power. Unintended side effects also get ever larger, with the climate crisis as the last culmination of an unintended side-effect of a new human technology. Superhuman AI may be the next technology on the list. Its side effects could make the climate crisis pale in comparison. Existential risk is the study of mostly unintended side-effects of our own ever more powerful technologies.

JWI: You've called for an international pause on AI development beyond ChatGPT 4. What specific risks do you see in allowing the current pace of AI research to continue unchecked?

Otto Barten: *The biggest risk I see is loss of control, which is also the risk the two most-cited AI professors are warning of, Yoshua Bengio (also Turing Award winner and a guest at our event earlier this year) and Geoffrey Hinton (also Turing Award and Nobel Prize winner). According to them, this may happen in five to twenty years, by which time we will direly need to come up with a solution. Also, a few hundred AI researchers have signed an open statement last year signalling they see a similar risk. Loss of control means that an AI gets so powerful, that it is beyond human means to stop it from executing any goal it may have. According to these experts, this may result in human extinction.*

A risk I am personally somewhat less concerned about but some others are, is losing control in a different way: not by a single AI during or shortly after development, but by many models that are applied in society. For example, the thinking goes, we could have AI CEOs, AI politicians, AI journalists, AI researchers, etc., which seem superficially aligned. However, later on, circumstances might change, and they would become unaligned. In a future world where most of the decision making

would be done by AIs, humanity would then not be in a position to resist newly unaligned AIs in powerful positions.

To make sure these risks do not materialise, we are calling for a Conditional AI Safety Treaty, which would entail that leading AI nations would pause training runs beyond a certain size, measured for example in flops [a measure of computer performance in operations per second], within their borders if and when capabilities are too close to loss of control level, to be judged by AI Safety Institutes.

JWI: In your opinion, why do market dynamics in AI development push companies to prioritise speed over safety? Can you point to any specific examples where this has already happened?

Otto Barten: *It seems obvious that market dynamics cause prioritising speed over safety. Market dynamics are by the way not the only mechanism causing this issue: academic pressure and political pressure by governments are two others. For a company, even if it is concerned about AI risks, it will be very tempting to release its models to make more money and not be left behind by competitors. One example where this happened is when tech companies quickly released LLMs [Large Language Models like ChatGPT] on the market after ChatGPT, even though they previously thought they were not fit for public use yet. It seems obvious that this will continue to happen. We think the only thing that will stop companies (and academics, and governments) from developing dangerous AI is international regulation.*

JWI: You propose an international moratorium on AI development. How realistic is it to get global players, especially countries like the U.S. and China, to agree on such a pause? Is it too late? Has an arms race already begun?

Otto Barten: *I do not believe it is too late. Arms races can be stopped. I believe, though, that it may be too early. The threat model that we are most concerned about and that warrants an international pause*

is loss of control. According to our research, only 15% of the US public is aware that AI might cause human extinction (up from 7% a few years earlier). As long as this percentage is at this relatively low level, I think it might be difficult to get an international treaty accepted. However, I believe this percentage will rise, fuelled by AI progress in the future.

Not all non-technologists understand that technology is reproducible and can therefore never get worse (according to some benchmark), but only better. It is a question of when, not if, AI will get better. Future AI progress will prompt more academics and others to voice existential risk concerns and this will increase public awareness. In the end, a treaty will likely get signed by the US, China, and other countries. The big question is whether this will happen soon enough (before loss of control). I don't have an answer to that question, but I think a treaty may well help, it may well be in time, and it is therefore very much worth trying.

JWI: Some AI researchers argue that pausing development might hinder important benefits that AI could bring. How do you respond to concerns that a pause might slow progress in beneficial areas such as medicine or climate change mitigation? How should we think about balancing these risks and benefits?

Otto Barten: *They are right, but avoiding human extinction is more important. Human extinction or a permanent loss of control would be a tragedy for us, everyone we love, our complete history and our complete future. Even something as important as finding new medicines or mitigating climate change is far less important than that. I think the public, and policymakers, will understand that.*

That having been said, it is one of the benefits of our Conditional AI Safety Treaty that it only kicks in when capabilities get too close to loss of control, and therefore does not slow progress until then. This makes sure we can get as many of the benefits, and as soon, as is responsible.

JWI: Do you see any particular significant risks for workers and work in the unchecked development of advanced artificial intelligence?

Otto Barten: *Yes, definitely. It is highly uncertain what, after the passing of an AI Treaty, is the level of technology that we can still responsibly use (meaning, what level is below the then mutually agreed loss of control red lines). If this permissible level of technology would be above what most white-collar workers currently achieve, I think they have reason to fear their professional futures. I also think that in such a scenario, the old maxim that workers should adapt and retrain, not resist, is not obviously valid as AI would be superhuman at many, perhaps all, formerly human occupations. I can see a future where everyone gets a universal basic income or, as some say, a universal high income and the work is mostly done by AI. That will require major societal adaptation but is thinkable.*

I can also see a future, though, where basically those without money will serve those with money forever, in jobs such as waiting their tables, raising their children, caring for them, entertaining them, etc., not because they are required but simply because the rich prefer to get served by poor people rather than AI-powered robots. This seems dystopian to me. In any case, I think it is important that workers fight for their share of power, their democratic rights, and their economic rights, and this will only become more important (and perhaps more challenging) with the advent of AI. One outcome of such fights could also be to prohibit AI application (even below the loss of control red lines), if we democratically choose to keep certain work human.

JWI: Could you elaborate on what threats you see advanced AI posing to decent jobs and people's ability to earn an income?

From the industrial revolution onwards, many forms of manual labour have been automated away and replaced by machines. What was left for humans, was thinking, which is why we saw an ever-increasing share of cognitive jobs in service industries. The AI revolution is now automating away thinking itself.

Most cognitive tasks can still be done better by humans than by AI, but it is the long-standing aim of the field of AI to end this, and create a situation where AI can outthink humans and take over most tasks. If AI would indeed become better at customer service, programming, driving, management, investment, sales, and basically all cognitive tasks, all these jobs will probably become superfluous.

JWI: Could you clarify the kinds of existential risks you are focused on? Is it potentially literally the death of humanity, or lesser catastrophic outcomes like mass job loss, job quality decline, significant economic upheaval or exacerbated inequalities or how are you defining this? I suppose my question is both about how you are defining and monitoring existential risk and what possible scenarios you envisage are likely enough outcomes to warrant preventative action through an AI safety treaty.

Otto Barten: *We're using Toby Ord's definition of existential risk, which is either human extinction, or a permanent dystopia, or an unrecoverable collapse. We are not an organisation focusing on all effects of AI, of which there are many. Non-existential effects of AI are more temporary by definition. I am personally more hopeful that democratic societies can mitigate these effects should they get a lot worse.*

For example, if inequality is exacerbated and e.g. 90% of the population would become poor, they can vote for redistribution. In principle, current democracies should be able to cope with such a challenge. Still, history has shown that optimal outcomes are by no means guaranteed, even in democracies. This is why organisations such as the Australian Institute of Employment Rights are so important.

JWI: Could you elaborate on the point at which you think Advanced AI development should go no further, i.e. elaborate more on what you mean by advanced or superintelligent A.I. and what some of the 'red lines' of control loss might be?

Otto Barten: *If and when AI gets so advanced that humanity is no longer able to stop it from achieving its goals, AI would be at loss of control-level. Unfortunately, it seems unlikely that we will know exactly where loss of control-level is. This means we will probably need to keep a large safety margin in between the maximum allowed level of AI capabilities and loss of control-level. The beginning of this safety margin is what I would refer to as our red lines.*

It is a current topic of research where these red lines should be. At the Existential Risk Observatory, we think capabilities such as being agentic, human manipulation, a good world model, situational awareness, long-term planning, (bio)weapons manufacturing, hacking, and self-improvement are likely relevant. Certain combinations of these capabilities are plausibly existentially dangerous. The UK AI Safety Institute includes "the ability for these systems to create copies of themselves online, to persuade or deceive humans, and to create more capable AI models than themselves." AI is already good at some of these capabilities, such as human manipulation. It is currently not good yet at others, such as long-term planning. Surprisingly little research has been done into linking threat models, including existential ones, to capabilities levels. We think significant progress can be made here and we recommend further research.

JWI: What role do you envision governments playing in enforcing AI safety and governance, particularly in industries driven by private tech companies?

Otto Barten: *I think they have an enormous role to play. Governments are, in my opinion, the only candidates to pass and enforce legislation that will keep us from becoming extinct, and that will keep other negative AI effects in check (such as the EU AI Act aims to do). I don't think governments should be the ones developing frontier AI themselves, as some propose. Private tech companies can develop the technology and aim to align or control it. Governments should make sure, by using democratically*

decided hard law and international treaties, that tech companies don't cross loss of control red lines, and that their products are not otherwise causing more problems than they solve.

JWI: Do you think existing regulatory frameworks, like those for nuclear weapons or climate agreements, offer useful models for AI governance, or does AI present unique challenges that require new approaches?

Otto Barten: *AI is definitely unique, but we can still learn from the past to an extent. In nuclear weapons, the International Atomic Energy Agency, the UN, non-proliferation treaties, and weapon reduction treaties have, to an extent, been successful in reducing nuclear proliferation. We never had a full-scale nuclear war, a fact we should appreciate a lot more. Still, we will need to do better with AI, since we cannot fail once in non-proliferation. Also, in AI, it is much less clear where the red lines are, and the equipment, such as training GPUs [Graphic Processor Chips used to train artificial intelligence models], may be more difficult to track than atomic equipment, especially in the future. These are mostly novel challenges that we need to work hard on today, to reduce our extinction chance possibly 5 to 20 years from now.*

JWI: Looking ahead, if a pause in AI development were successfully implemented, what steps would need to be taken during that time to ensure that AI research can proceed safely in the future?

Otto Barten: *I am not sure we will ever be able to responsibly build AI beyond loss of control level. I am sure though that companies will try to continue development. If they meet certain democratically agreed-upon criteria, perhaps they could continue. For example, nuclear reactor approval requires quantifying the meltdown risk at below 0.0001% per year. If AI companies could somehow guarantee that we will not lose control over their technology with a similar likelihood, they could perhaps continue*

development. I am not an expert in how they should achieve this. I think it will be very difficult, but we have succeeded in solving many very difficult problems, so I don't rule out the option.

Declaration of interests

Nil.

James Fleming
October 2024

