AI Risk for Epistemic Minimalists

Alex Flint

August 22, 2021

Financial status: This is independent research, now supported by a grant. I welcome further financial support.

Epistemic status: This is an attempt to use only very robust arguments.

Outline

- I outline a case for concern about AI that does not invoke concepts of agency, goal-directedness, or consequential reasoning, does not hinge on single- or multi-principal or single or multi-agent assumptions, does not assume fast or slow take-off, and applies equally well to a world of emulated humans as to de-novo AI.
- The basic argument is about the power that humans will temporarily or permanently gain by developing AI systems, and the history of quick increases in human power.
- In the first section I give a case for paying attention to AI at all.
- In the second section I give a case for being concerned about AI.
- In the third section I argue that the business-as-usual trajectory of AI development is not satisfactory.
- In the fourth section I argue that there are things that can be done now.

The case for attention

We already have powerful systems that influence the future of life on the planet. The systems of finance, justice, government, and international cooperation are things that we humans have constructed. The specific design of these systems has influence over the future of life on the planet, meaning that there are small changes that could be made to these systems that would have an impact on the future of life on the planet much larger than the change itself. In this sense I will say that these systems are powerful.

Now every single powerful system that we have constructed up to now uses humans as a fundamental building-block. The justice system uses humans as judges and lawyers and administrators. At a mechanical level, the justice system would not execute its intended function without these building-block humans. If I turned up at a present-day court with a lawsuit expecting a summons to be served upon the opposing party but all the humans in the justice system were absent then the summons would not end up being served.

The Google search engine is a system that has some power. Like the justice system, it requires humans as building-blocks. Those human building-blocks maintain the software, data centers, power generators, and internet routers that underlie it. Although individual search queries can be answered without human intervention, the transitive closure of dependencies needed for the system to maintain its power includes a huge number of humans. Without those humans, the Google search engine, like the justice system, would stop functioning.

The human building-blocks within a system do not in general have any capacity to influence or shut down the system. Nor are the actions of a system necessarily connected to the interests of its human building-blocks.

There are some human-constructed systems in the world today that do not use humans as building-blocks, but none of them have power in their own right. The Curiosity Mars rover is a system that can perform a few basic functions without any human intervention, but if it has any influence over the future, it is via humans collecting and distributing the data captured by it. The Clock of the Long Now, if and when constructed, will keep time without humans as building-blocks, but, like the Mars rovers, will have influence over the future only via humans observing and discussing it.

Yet we may soon build systems that do influence the future of life on the planet, and do not require humans as building-blocks. The field concerned with building such systems is called artificial intelligence and the current leading method of engineering is called machine learning. There is much debate about what exactly these systems will look like, and in what way they might pose dangers to us. But before taking any view about whether these systems will look like agents or tools or AI services, or whether they will be goal-directed or influence-seeking, or whether they will be developed quickly or slowly, or whether we will end up with one powerful system or many, we might ask: what is the least we need to believe to justify attending to the development of AI among all the possible things that we might attend to? And my sense is just this: we may soon transition from a world where all systems that have power over the future of life on the planet are intricately tied to human building-blocks to a world where there are some systems that have power over the future of life on the planet without relying on human building-blocks. This alone, in my view, justifies attention in this area, and it does not rest in any way on views about agency or goals or intelligence.

So here is the argument up to this point:

> Among everything in the world that we might pay attention to, it makes sense to attend to that which has the greatest power over the future of life on the planet. Today, the systems that have power over the future of life on the planet rely on humans as building-blocks. Yet soon we may construct systems that have power but do not rely on humans as building-blocks. Due to the significance of this shift we should attend to the development of AI and check whether there is any cause for concern, and, if so, whether those concerns are already being adequately addressed, and if not, whether there is anything we can do.

The case for concern

So we have a case for paying some attention to AI among all the things we could pay attention to, but I have not yet made a case for being *concerned* about AI development. So far it is as if we discovered an object in the solar system with a shape and motion quite unlike a planet or moon or comet. This would justify some attention by humans, but on this evidence alone it would not become a top concern, much less a top cause area.

So how do we get from attention to concern? Well, the thing about power is that *humans already seek it*. In the past, when it has become technically feasible to build a certain kind of system that exerts influence over the future, humans have tended, by default, to eventually deploy such systems in service of their individual or collective goals. There are some classes of powerful systems that we have coordinated to avoid deploying, and if we do this for AI then so much the better, but by default we ought to expect that once it becomes possible to construct a certain class of powerful system, humans will deploy such systems in service of their goals.

Beyond that, humans are quite good at incrementally improving things that we can tinker with. We have made incremental improvements to airplanes, clothing, cookware, plumbing, and cell phones. We have not made incremental improvements to human

minds because we have not had the capacity to tinker in a trial-and-error fashion. Since all powerful systems in the world today use humans as building blocks, and since we do not presently have the capacity to make incremental improvements to human minds, there are no powerful systems in the world today that are subject to incremental improvement at all levels.

In a world containing some powerful systems that do not use humans as building blocks, there will be *some powerful systems that are subject to incremental improvements at all levels*. In fact the development of AI may open the door to making incremental improvements to human minds too. In this case *all* powerful systems in the world would be subject to incremental improvement. But we do not need to take a stance on whether this will happen or not. In either case the situation we will be in is one in which humans are making incremental improvements to some systems that have power in the world, and we therefore ought to expert that the power of these systems will therefore increase on a timescale of years or decades.

Now at this point it is sometimes argued that a transition of power from humans to non-human systems will take place, due to the very high degree of power that these non-human systems will eventually have, and due to the difficulty of the alignment problem. But I do not think that any such argumentative move is necessary to justify concern, because whether humans eventually lose power or not, what is much more certain is that in a world where powerful systems are being incrementally improved, there will be a period during which *humans gain power quickly*. It might be that humans gain power for mere minutes before losing it to a recursively self-improving singleton, or it may be that humans gain power for decades before losing it to an inscrutable web of AI services, or it may be that humans gain power and hold onto it until the end of the cosmos. But under any of these scenarios, humans seem destined to gain power on a timescale of years or decades, which is the pace at which we usually make incremental improvements to things.

What happens when humans gain power? Well as of today, existential risk exists. It would not exist if humans had not gained power over the past few millennia, or at least it would be vastly reduced. Let's ignore existential risk due to AI in order to make sure the argument is non-circular. Still, the point goes through. There is much good to say about humans. This is not a moral assessment of humanity. But can anyone deny that humans have gained power over the past few millennia, and that, as a result of that, existential risk is much increased today compared to a few millennia ago? If humans *quickly gain power*, it seems that, by default, we ought to presume that existential risk will also increase.

Now, there are certainly *some ways* to increase human power quickly without increasing existential risk, including by skillful AI development. There have certainly been *some times and places* where rapid increases in human power have led to decreases in existential risk. But this part of the argument is about what happens by default, and the ten thousand year trendline of the "existential risk versus human power" graph is very much up-and-to-the-right. Therefore I think rapidly increasing human power will increase existential risk. We do not need to take a stance on how or whether humans might later lose power in order for this to go through. We merely need to see that, among all the complicated goings-on in the world today, the development of AI is the thing most likely to confer a rapid increase in power to humans, and on the barest historical precedent, that is already cause for both attention and concern.

So here is the case for concern:

> If humans learn to build systems that do influence the future of life on the planet but do not require human building-blocks, then they are likely to make incremental improvements to these systems over a timescale of years or decades, and thereby increase their power over the future of life on the planet on a similar timescale. This should concern us because quick increases in human power have historically led to increases in existential risk. We should therefore investigate whether these concerns are already being adequately

addressed, and, if not, whether there is anything we can do.

I must stress that not all ways of increasing human power lead to increases in existential risk. It is as if we were considering giving a teenager more power over their own life. Suppose we suddenly gave this teenager the power not just of vast wealth and social influence, but also the capacity to remake the physical world around them as they saw fit. For typical teenagers under typical circumstances, this would not go well. The outcomes would not likely be in the teenager's own best interests, much less the best interests of all life on the planet. Yet there probably *are* ways of conferring such power to this teenager, say by doing it slowly and in proportion to increases in the teenager's growing wisdom, or by giving the teenager a wise genie that knows what is in the teenager's best interest and will not do otherwise. In the case of AI development, we are collectively the teenager, and we must find the wisdom to see that we are not well-served by rapid increases in our own power.

The case for intervention

We have a case for apriori concern about the development of a particular technology that may, for a time, greatly increase human power. But perhaps humanity is already taking adequate precautions, in which case marginal investment might be of greater benefit in some other area. What is the epistemically minimal case that humanity is not already on track to mitigate the dangers of developing systems that have power over the future of life on the planet without requiring humans as building-blocks?

Well consider: right now we appear to be rolling out machine learning systems at a rate that is governed by economic incentives, which is to say that the rate of machine learning rollout appears to be determined primarily by the supply of the various factors of production, and the demand for machine learning systems. There is seemingly no gap between the rate at which we *could* roll out machine learning systems if we allowed ordinary economic incentives to govern, and the rate at which we *are* rolling out those systems.

So is it more likely that humanity is exercising diligence and coordinated restraint in the rollout of machine learning systems, or is it more likely that we are proceeding haphazardly? Well imagine if we were rolling out nuclear weapons at a rate determined by ordinary economic incentives. From a position of ignorance, it's *possible* that this rate of rollout would have been selected by a coordinated humanity as the wisest among all possible rates of rollout. But it's much more likely that this rate is the result of haphazard discoordination, since from economic arguments we would expect the rate of rollout of any technology to be governed by economic incentives in the absence of a coordinated effort, whereas there is no reason to expect a coordinated consideration of the wisest possible rate to settle on this particular rate.

Now, if there were a gap between the "economic default" rate of rollout of machine learning systems and the actual rate of rollout then might still question whether we were on track for a safe and beneficial transition to a world containing systems that influence the future of life on the planet without requiring humans as building-blocks. It might be that we have merely placed haphazard regulation on top of haphazard AI development. So the existence of a gap is not a sufficient condition for satisfaction with the world's handling of AI development. But the absence of any such gap does appear to be evidence of the absence of a well-coordinated civilization-level effort to select the wisest possible rate of rollout.

This suggests that the concerning situation in the previous section is, at a minimum, not already *completely* addressed by our civilization. It remains to be seen whether there is anything we can do about it. The argument here is about whether the present situation is already satisfactory or not.

So here is the argument for intervention:

> Humans are developing systems that appear destined to quickly increase human power over the future of life on the planet at a rate that is consistent

> with an economic equilibrium. This suggests that human civilization lacks the capacity to coordinate on a rate motivated by safety and long-term benefit. While other kinds of interventions may be taking place, the absence of this particular capacity suggests that there is room to help. We should therefore check whether there is anything that can be done.

Now it may be that there is a coordinated civilization-level effort that is taking measures other than selecting a rate of machine learning rollout that is different from the economic equilibrium. Yes, this is possible. But the question is why our civilization is not coordinating around a different rate of machine learning rollout if it has the capacity to do so. Is it that the economic equilibrium is in fact the wisest possible rate? Why would that be? Or is it that our civilization is choosing not to select the wisest possible rate? Why? The best explanation seems to be that our civilization does not presently have an understanding of which rates of machine learning rollout are most beneficial, or the capacity to coordinate around a selected rate.

It may also be that we navigate the development of powerful systems that do not require humans as building-blocks without ever coordinating around a rate of rollout different from the economic equilibrium. Yes this is possible, but the question we are asking here is whether humanity is already on track to safely navigate the development of powerful systems that do not require humans as building-blocks, and whether our efforts would therefore be better utilized elsewhere. The absence of the capacity to coordinate around a rate of rollout suggests that there is at least one very important civilizational capacity that we might help develop.

The case for action

Finally, the most difficult question of all: is there anything that can be done? I don't have much to say here other than the following very general point: It is very strong to claim that nothing can be done about a thing because there are many possible courses of action, and if even one of them is even a little bit effective then there is something that can be done. To rule out all possible courses of action requires a very thorough understanding of the governing dynamics of a situation and a watertight impossibility argument. Perhaps there is nothing that can be done, for example, about the heat death of the universe. We have some understanding of physics and we have strong arguments from thermodynamics, and even on this matter there is some room for doubt. We have nowhere near that level of understanding about the dynamics of AI development, and therefore we should expect on priors that among all the possible courses of actions, there are some that are effective.

Now you may doubt whether it is possible to *find* an effective course of action. But again, claiming that it is impossible to find an effective course of action implies that among all the ways that you might try to find an effective course of action, none of them will succeed. This is the same impossibility claim as before, only now it concerns the process of finding an effective course of action rather than the process of averting AI risk. Once again it is a very strong claim that requires a very strong argument, since if even one way of searching for an effective course of action would succeed, then it is possible to find an effective course of action.

Now you may doubt that it is possible to find a way to search for an effective course of action. Around and around we could go with this. Each time you express doubt I would point out that it is not justified by anything that is objectively impossible. What, then, is the real cause of your doubt?

One thing that can always be done at an individual level is to make a thing the top priority in our life, and to become willing to let go of all else in service of it. At least then if a viable course of action does become apparent, we will certainly be willing to take it.

Conclusion

In the early days of AI alignment there was much discussion about fast versus slow take-off, and about recursive self-improvement in particular. Then we saw that *the situation is concerning either way*, so we stopped predicating our arguments on fast take-off, not because we concluded that fast take-off arguments were wrong, but because we saw that the center of the issue lay elsewhere.

Today there is much discussion in the alignment community about goal-directedness and agency. I think that a thorough understanding of these issues is central to a solution to the alignment problem, but, like recursive self-improvement, I do not think it is central to the problem itself. I therefore expect discussions of goal-directedness and agency to go the way of fast take-off: not dismissed as wrong, but de-emphasized as an unnecessary predicate.

There is also discussion recently about scenarios involving single versus multiple AI systems governed by single versus multiple principals. Andrew Critch has argued that more attention is warranted to "multi/multi" scenarios in which multiple principals govern multiple powerful AI systems. Amongst the rapidly branching tree of possible scenarios it is easy to doubt whether one has adequately accounted for the premises needed to get to a particular node. It may therefore be helpful to lay out the part of the argument that applies to all branches, in order that we have some epistemic ground to stand on as we explore more nuance. I hope this post helps in this regard.

Appendix: Agents versus institutions

One of the ways that we could build systems that have power over the future of life on the planet without relying on human building-blocks is by building goal-directed systems. Perhaps such goal-directed systems would resemble agents, and we would interact with them as intelligent entities, as Richard Ngo describes in AGI Safety from First Principles.

A different way that we could build systems that have power over the future of life on the planet without relying on human building-blocks is by gradually automating factories, government bureaucracies, financial systems, and eventually justice systems as Andrew Critch describes. In this world we are not so much interacting with AI as a second species but more as the institutional and economic water in which we humans swim, in the same way that we don't think of the present-day finance or justice systems as agents, but more like a container in which agents interact.

Or perhaps the first systems that will have power over the future of life on the planet without relying on human building-blocks will be emulations of human minds, as Robin Hanson describes in Age of Em. In this case, too, humans would gain the capacity to tinker with all parts of some systems that have power over the future of life on the planet, and through ordinary incremental improvement become, for a time, extremely powerful.

These possibilities are united as avenues by which humans could quickly increase their power by building systems that have both influence over the future of life on the planet, and are subject to incremental improvement at all levels. Each scenario suggests particular ways that humans might later lose power, but instead of taking a strong view on the loss of power we can see that a quick increase in human power, however temporary, is, on historical precedent, already a cause for concern.