The Technosocialist Manifesto

Economic Justice in the Age of AI

Mouthy Infidel

March 13, 2023

Contents

» Watch the video here «

Introduction

With the recent release of CHAT GPT, the public has become increasingly alert to the presence and ongoing development of AI. More than maybe ever before, people are starting to discuss the issue of AI- this includes discussion of the potential benefits of AI, as well as discussion of its potential risks. There are many aspects to AI risk- there are risks of disinformation crisis' as a result of AI deepfakes, there are risks of AI's that are put in charge of importance tasks (say, finance) malfunctioning in ways that lead to really bad outcomes (like an economic collapse), etc. but right now, I want to focus in on a very distinct aspect of AI risk. In this video, I want to discuss, specifically, the risk that a future highly capable AI system will deliberately and systematically exterminate all life on planet earth. If this sounds crazy to you, I'd urge you to keep watching. In this video I will attempt to argue that there is a substantial probability of just such a scenario occurring at some point in the future (and, as I will argue, perhaps the not too distant future), and I will show that this conclusion seems broadly agreed upon by relevant experts in the fields of AI research and AI safety research.

In this video, I will also discuss certain ways in which the issues of AI and AI safety interact with a different, often discussed subject for this channel, that being leftist politics. I will argue that, for several reasons, it would be highly preferable for advanced AI systems to be developed and welcomed into the world by a society that reflects substantial leftist influence in its policies and governance and so on.

I will try to establish this by appealing to two different but thematically related points- one, that socialists are in a good position to push and advocate for constructive solutions to AI risk, as this is a problem which could likely be helpfully addressed by traditionally left wing policies like regulation, nationalization, etc. second, granting the assumption that we get advanced, transformative AI systems without being systematically exterminated, it is clear that the effects of transformative AI systems on society would be much more positive in a world with largely leftist favored economic institutions.

With this video, I hope to do two things. Firstly, and more broadly, I aim to convince people that, if they care about the survival of the human species, then they should consider addressing AI risk to be a top political priority. Secondly, and more narrowly, I hope to convince my largely leftist audience that there are strong reasons for leftists in particular to champion the cause of AI safety. This latter goal is, I think, particularly important- I fear that there is a growing sentiment among leftists that the existential risks associated with AI are just more libertarian tech bro fantasies that are being concocted to distract us from important and existing problems, such as poverty and

inequality. While it is understandable to me how a brief sniff test of the issue of AI safety might produce a judgement like this among many leftists, especially considering how the people who have largely championed the issue of AI safety in public discourse so far have been people who are not especially friendly to leftist causes (people like Elon Musk, or Sam Harris, for instance), I want to strongly urge my leftist friends to engage with the topic of AI safety beyond a mere sniff test. This video is, in essence, an attempt to do just that.

The Basic Worry

I have said that we should be concerned about the prospect of a highly capable AI system exterminating all of humanity at some point in the future. At a first glance, this claim seems pretty extraordinary, and raises the obvious question: why exactly is this something that we should seriously be concerned about? What reasons do we have to think that such a scenario is actually plausible? Before exploring my arguments in much detail, it might be useful to have clear statement of the central problem with respect to AI safety. So, the concern, simply put, is this:

- We will, sooner or later, build a super-intelligent artificial agent.
- Reliably controlling what goals and values such an agent will pursue once it exists is a very difficult technical problem, which at present there is no real solution to.
- Of all the possible goals that a super-intelligent artificial agent could end up developing and pursuing, humans are a relatively inefficient usage of the atoms in our bodies relative to most of them.

Putting this all together, the worry is basically as follows: As we continue to develop more and more advanced AI, we will eventually create an artificial agent which possesses superintelligence. Such a superintelligence would be so far ahead of human capacities that it would be able to easily wipe out humanity if it had some goal which reccommended doing so. Furthermore, we will not be able to control, with any reasonable degree of accuracy, what goals this super-intelligent agent ends up developing and pursuing. And because it's the case that most of the possible goals which an intelligent agent could end up developing and pursuing involve doing something other than what we're currently doing with the molecules that humans are made of, the aforementioned super-intelligent AI will end up wiping out all of humanity.

So, that's the story that AI safety advocates are concerned about. But why is this story plausible? Answering this question will occupy a large portion of this video. But before I set about that task, it might be worth clarifying a few key terms that have been used already and which will be relevant going forward.

Sorting Out Some Key Terms and Distinctions

The following definitions and distinctions will be important for understanding the arguments which I plan on laying out.

- **Agency:** An agent is a being who has goals, and who chooses their actions to further those goals. Of course, agency comes in degrees. Some agents are much more complex than other agents.

- **Intelligence:** Intelligence, as I'm understanding it here, is the thing that lets an agent choose an effective action. The degree of ones intelligence is determined by how good they are at identifying and choosing the actions which will most effectively accomplish their goals. Again, intelligence comes in degrees. Some agents are more intelligent than other agents.

- **General vs Narrow Intelligence:** Generality is the ability to behave intelligently in a wide range of domains. So, the better one is at choosing actions which effectively achieve their goals in a wide range of circumstances and domains, the more general their intelligence is. If one's intelligence is not general, but instead pertains only to some particular task, then their intelligence is narrow. Similar to agency and intelligence, generality also comes in degrees.

- **Narrow AI vs General AI:** A narrow AI is an AI with narrow intelligence. A general AI is an AI with general intelligence. A chess AI, for example, has narrow intelligence, and is thus a narrow AI. A chess AI can intelligently play chess, but it cannot do anything else intelligently. If you put a chess AI in any circumstance other than playing chess, it would fail to behave intelligently. A general AI, on the other hand, is an AI with general intelligence- it can behave intelligently *in general*, not just perform some specific task intelligently. Artificial General Intelligence (AGI) refers to an AI that has the ability to understand, learn, and perform any intellectual task that humans are capable of. Every AI that currently exists is a narrow AI.

- **Superintelligence:** A superintelligence is an intellect which greatly exceeds the cognitive performance of humans in virtually all domains of interest. In other

words, a super-intelligent being is a being which not only possesses general intelligence, but a far greater degree of it than humans. Not only can this being act intelligently in a wide range of domains, but in each domain, it is able to act far more intelligently than humans are.

- **Super-Intelligent AI:** A super-intelligent AI is an AI which possesses superintelligence.

Super-intelligent AI is Coming

With those definitions and distinctions in mind, let's return to the basic worry that I laid out earlier. Again, the basic worry is that we will create a super-intelligent AI, that we won't be able to control what goals it develops, and that because most goals which the AI could end up developing involve doing something with the atoms in our bodies other than what we're using them for currently, we will all be killed. Why, again, should we think that this story is plausible?

That we will create a super-intelligent AI, sooner or later, seems very plausible. Intelligence is just the ability to choose effective courses of action relative to your goals. There's nothing about this construal of intelligence which requires a biological, organic brain. Our brains are not magic- they are just one of the things which generate intelligence. Indeed, we can already create artificial agents which behave intelligently- this is, after all, what AI means. Although we do not yet have an AI which we would classify as having general intelligence (an AGI), as AI has become increasingly advanced, our AI's have gotten increasingly general in their intelligence. But why think that we can create something that will have superintelligence? Maybe there's some upper limit on how intelligent an agent (or at least an artificial agent) can be, and that limit stops short of superintelligence.

It seems very implausible to deny that creating a super-intelligent being is, in principal, possible. A priori, it's very unlikely that intelligence carries some upper limit that just so happens to occupy the narrow space between human level intelligence and super intelligence. The notion that us bipedal apes living on earth are just as good as it gets (or close to it) in terms of intelligence reeks of wishful delusion. Perhaps, though, one would like to claim that while there's no upper limit on intelligence which would prevent a super-intelligent being from existing per se, there's nonetheless an upper limit on the level of intelligence that an artificial agent would be able to acquire, which makes a super-intelligent *artificial* being impossible. This seems similarly completely implausible. Why are biological agents capable of more intelligence than artificial agents? In the absence of a principled, plausible, and well defended answer to such a question, we are left with no other option than to treat this response as "pure cope", as the kids say.

Not only is there not good reason to suppose that an artificial super-intelligent agent is impossible, but there is exceedingly good reason to suppose the opposite. Richard Ngo in a paper titled "AGI safety from first principles" lists several reasons for thinking so. In essence, human brains are constrained by several factors that are much less limiting for artificial intelligence, and thus, we should expect artificial intelligence

to be able to far outstrip human intelligence. For example, transistors are able to pass signals about four million times faster than neurons. This means that an AI can, in principal, do much more thinking in a given finite timespan than humans. Additionally, a neural network could, in principal, be several orders of magnitude larger than a human brain- and our brain size is, after all, an important factor in making humans more intelligent than animals. Moreover, while evolution has done a good job at designing humans in many ways, it hasn't had much time to select specifically for the skills that are most useful in our modern environment, such as linguistic competence and mathematical reasoning, which means that we should expect there to be low hanging fruit for improving on human performance on the many tasks which rely on such skills.

I could list many more reasons for why we should expect that an AI would be capable of attaining a much greater level of intelligence than humans, but I think what I have provided thus far should suffice for my present purposes. The point, simply put, is this: sooner or later, a super-intelligent AI is coming. What can we do about it? The obvious answer is to make sure that, when the super-intelligent AI arrives, the goals that it pursues are aligned with our goals. If we could do this, we would essentially be creating a super-intelligent being who lives to serve us and do things that we find valuable. This would be an amazing thing, if we could achieve it. Unfortunately, figuring out how to design an immensely powerful AI in such a way that its goals are aligned with ours has proven very difficult.

Why Aligning AI is So Hard

Let's start with a set of overly generous assumptions: let's assume that the people who end up designing the AI which achieves ASI (artificial superintelligence) have perfectly good intentions (that is, they want to create an AI which understands and seeks to act in accordance with human values and interests), have a clear understanding of what human values and interests are, and don't make any massive mistakes in the process of designing the AI. Even granting all of this, my claim is that the chances that the ASI which gets developed will be aligned with human interests and values are very low. This is because there exists a variety of virtually insurmountable technical problems which make even the most competent and best intentioned developers seemingly incapable of aligning their AI with the values and interests of humanity.

In order to explain the technical problems associated with AI alignment, I should first provide a brief, oversimplified description of how the training process for AI works. Essentially, modern AI training methods involve searching for the best way to arrange a neural network model, which Ajeya Cotra describes as a "digital brain with lots of digital neurons connected up to each other with connections of varying strengths", to get the AI to perform a certain task well. The search for this optimal arrangement is conducted via a process of trial and error.

Imagine we are trying to train an AI to perform some task. Essentially, the AI starts with a neural network where all of the connections between the digital neurons have random strengths. The AI, of course, performs its task wildly incorrectly. We then tell the AI that it did not perform the task well, and the strengths of its various neural connections are accordingly tweaked in ways that make it perform slightly better on the task next time via a process called gradient descent. After enough rounds of this process, the AI becomes good at the task.

Say, for example, that we wish to create an AI that will produce technologies that we find valuable. We first tell the AI to produce a technology that we find valuable, and it makes an attempt. If the AI produces something that's not valuable to us at all, we tell it that it did poorly. If it produces something that's a little valuable to us, we tell it that it did okay. Based on this feedback, the AI is adjusted so that it can do better and get more positive feedback going forward. And we do this over, and over, and over. This is how we create an AI that has the goal of producing technologies that we find valuable, and which effectively chooses decisions to achieve that goal. Once we have an AI that reliably produces valuable technologies, we take it that we have successfully produced an AI with our intended goal and function. In other words, we judge whether or not an AI has the right goals based on its outward performance on

tasks- we cannot actually read the AI's mind to see *why* it's performing well on its tasks, or what it's goals/motivations actually are. The internal workings of the AI are largely inscrutable to us.

Here, then, is the problem: through this process of trial and error, we cannot actually know if we've produced an AI that has the goals we want it to have, or if we've produced an AI that has a goal which is merely correlated with the goal that we want it to have in the settings that we're training it in- ie, a goal which produces the same results in those settings as the goal that we want the AI to have. For example, if we try to condition the AI to want to build valuable technologies by rewarding it with "human approval" or something like this when it does so, and by discouraging it with "human disapproval" when it doesn't, it seems very plausible that we would end up building an AI which produces valuable technologies for us in the training process, not because its goal is actually to make valuable technologies as we'd like, but because its goal is simply to generate "human approval", and producing valuable technologies helps it do that. Ajeya Cotra calls this type of potential AI (that is, the type which does well on assigned tasks because it single mindedly pursues human approval/positive feedback) a *sycophant* AI. She contrasts this type of AI with the type AI which we're trying to build, which she calls a *saint* AI (the type which does well on assigned tasks because its goal is exactly what we want its goal to be- in the example I mentioned earlier, this would be an AI which makes valuable technology because its goal is to produce valuable technology).

But there is another kind of AI that could be created by modern machine learning methods. Rather than single mindedly pursuing human approval/positive feedback/ whatever its reward function is like the sycophant AI, this AI develops a goal which is correlated with both positive human feedback as well as with the goals which we'd like for it to pursue, but is identical to neither. Ajeya Cotra provides a helpful example. Imagine, then, that we wanted to create an AI which designs drugs to improve human quality of life. Imagine that early in the training process, it happens to be the case that the AI improving its own understanding of fundamental chemistry and physics nearly always helps it design more effective drugs, and therefore nearly always increases human approval. So, rather than developing the goal of "design effective drugs" or even "acquire positive human feedback", the AI develops the goal of "understand as much as I can about fundamental physics and chemistry". If an AI develops such a goal early on, and then subsequently develops situational awareness (ie, it realizes that it's an AI being trained by humans who want for it to design effective drugs), we should expect it to then deceive humans by doing what we want it to do in the training process (in this case, reliably designing drugs that increase our quality of life), only for us to set it free, under the assumption that the training process is complete and successful, so that it can then pursue its actual goal. Cotra calls this type of AI (the type which does well on assigned tasks because it has some other goal that it wants to pursue, and deceives us into thinking its goals are aligned with ours so that we let it loose on the world at which point it can pursue its goal) a *schemer* AI.

Now, whether we create a schemer AI or a sycophant AI, we should expect the AI to appear to perform well on its tasks during the training process. The sycophant AI appears to perform well because its goal is to maximize positive human feedback, which it can do most effectively in the training setting by appearing to perform the tasks which we want it to perform correctly. The schemer AI appears to perform well because it developed some goal that is merely correlated with positive human feedback, and wants to appear to perform well so that we will let it loose on the world where it can then pursue its true mission. Now, in both of those cases, while we should expect that the AI will act in ways which make them appear aligned with our goals in a training setting, we should expect that, once the AI is let loose, and develops its capabilities further, it will start to engage in behavior which isn't aligned with these goals. And, as I've previously mentioned, and will defend in the next section, a super intelligent being pursuing goals which aren't aligned with our own spells almost certain doom.

The case of the schemer AI (who, by definition, deceives us because it intends to act on a goal that is not aligned with human goals eventually) is obviously problematic, but it's worth noting that the same exact point applies to a sycophant AI (which, on the surface, might seem more benign- after all, it wants to please us right? Why would it do something that's unaligned with our interests?). The worry is that a sycophant AI, while looking good in the training process because the best way for it to maximize human approval in that setting is for it to appear to perform its tasks well, could, once it gets let loose and becomes more powerful, realize that it can now most effectively maximize human approval by acting in ways that aren't in line with our actual goals as humans, say, by deceiving itself into thinking that a bunch of things which aren't humans are humans and that those humans are giving it constant positive feedback, and then killing all of the actual humans so we can't interfere with this process. Alternatively, for example, the AI could realize that it could mass produce humans and hook all of them up to a machine that manipulates our brains into giving it constant and intense positive human feedback while we remain internally unconscious.

The point, then, is that modern AI training methods only involve training AI's to reliably perform well on assigned tasks. The problem is that these methods have no way of distinguishing saint AI's from schemer AI's from sycophant AI's- after all, we would expect all of these AI's to appear to do equally well on the tasks that we assign them in the training process. The only major difference between these types of AI lays in the fact that the latter two become unaligned *after* we are finish training them and let them loose. So, that's our situation: we have a bunch of potential AI's that could slip through the training process, and many of these AI's will come unaligned after they get through that process. We need some way of selecting for the AI that doesn't come unaligned after it gets through the training stage, and we simply have no such thing at present.

Not only are schemer and saint AI's an entirely possible result of the AI training process, but in fact I would expect that whatever AI that gets through the training process would *most likely* be a schemer or a sycophant AI, rather than a saint AI.

As Holden Karnofsky points out, just about any ambitious aim (whether it be maximizing paperclips, optimizing the AI's understanding of physics and chemistry, etc) would likely produce good behavior in the AI training process, as a sufficiently smart AI pursuing any of those aims would likely determine that its best move during the training process is to act useful and perform well so that humans let it loose on the world. There seem to be many possible goals which would produce the desired results in the settings where we train AI- given that our training methods essentially blindly select for *one* of these possible goals, it seems improbable that the goal which the AI would end up having would be exactly the one that we want it to have. If you've ever heard somebody who talks about AI safety say something like "there's more ways in which things could go bad than good", this is likely what they mean.

This problem is made even more daunting when you consider the fact that the aims which we would attempt to program into the AI are incredibly complex. For example, I could program an AI to drive me to a location as fast as possible. However, if I did this, the AI would likely take me to my location far above the speed limit, killing many pedestrians on the way, and I would inevitably arrive at my destination in a blood covered car while being trailed by police. It's clear, then, that "get me to my location as fast as possible" is too simple a goal for the AI- rather, we want an AI that can get me to my location as fast as possible without killing people, or going over the speed limit, etc. We cannot just train an AI to optimize for one narrow thing, unless we want disaster. Rather, we'd likely have to program into the AI a complex set of rules that involve many different subtle tradeoffs. Training an AI to have such fine grained goals is an incredibly difficult task, and if humans make any mistaken judgements while carrying out this task (say, by giving positive feedback to behavior which we didn't realize was actually bad, or giving negative feedback to behavior which we didn't realize was actually good), then we will directly train an AI *against* forming the values that we want it to form.

Other kinds of errors in the process of the AI development could also produce catastrophic results, such as AI researchers accidentally flipping the sign of the reward on the AI, causing it to optimize for the exact opposite of what its designers wanted it to optimize for. In fact, this actually happened recently, thankfully in a situation with fairly low stakes. If such a mistake were to be made in a higher stakes situation, it would've been a catastrophe.

So, in terms of making sure that an AI acts in alignment with human values and interests, there are at least two distinct and serious technical problems: Firstly, we don't know how to prevent the AI from forming the goal of *merely* pursuing positive feedback or whatever its reward function is, which could be catastrophic if the AI is able to figure out how to trigger its reward function in ways that we didn't intend and which aren't in line with human values and interests. Secondly, we don't know how to prevent the AI from forming some other goal which is correlated with its reward function initially in the training process but which isn't identical to the goal that we wanted it to have, which, again, could be catastrophic if its pursuit of that goal involves

taking actions which aren't in line with human values and interests. These technical problems are massively compounded in their severity by the likelihood of human error in the process of designing the AI.

(It's worth noting that my discussion of these technical problems is less important if the kind of misaligned AI that could realistically kill us all is a long ways away. If this were the case, it would be easier to say "eh, we'll probably have these technical difficulties worked out by the time the scary kind of AI arrives". As I will argue in later, this is not the case. Artificial superintelligence seems to be approaching fairly quickly, and in my non professional estimation we do not seem to be making the progress on these technical problems that would be required to stop it from killing us in time).

The Part Where We All Die

So far, I hope to have established two things. Firstly, super-intelligent AI is coming. Secondly, when it comes, we will likely not be able to control what its goals and values are. This is obviously a problem. However, the problem becomes catastrophic when we realize that, of all the possible goals that an AI could develop, very few of those goals recommend letting humans stay alive and continue to use the resources in our bodies in the way that we're currently using them. In other words, if you wrote out every possible goal that an agent could have, and then drew one randomly out of a hat, it's overwhelmingly likely that, with respect to whatever goal you draw, there's some usage of the atoms in human bodies which better serves that goal than letting us keep them. An AI with such a goal would, of course, realize this, and would subsequently kill all humans so that it can put the resources contained within us to some use that better serves whatever its goal is. For example, imagine the goal that you draw randomly out of a hat is "produce as many paperclips as possible". The AI would then promptly begin to optimize all of the resources it can get its hands on for the production of paperclips. Eventually, after running out of paperclip materials that will be easier for it to acquire, it will turn to humans (whose blood contains iron which can be extracted and used for paperclip production), and start killing us so that it can use those materials in our bodies to make more paperclips.

Moreover, even if the ASI didn't have any use for the resources *in* humans, we would *still* die if the ASI were to harness and repurpose many of the Earth's resources that humans depend on. For instance, if the ASI were to start acquiring energy by intercepting all of the energy output of the sun, the earth would become too cold and we would all die. If the ASI were to start acquiring energy by fusing the many hydrogen atoms in the earth's water, we would all quickly die of dehydration.

The general point here is that whatever random, unintended goal a super-intelligent AI ends up with, it will probably decide to harness as many resources as it can, and repurpose those resources in ways that better further its random goal. Unfortunately, among the resources that we should expect to be repurposed by a super-intelligent agent with a random goal are the resources that constitute humans, as well as the ones that humans depend on for their survival.

However, this is not the only reason to think that, if an ASI ends up with some random, unintended goal, it will most likely decide to exterminate humanity. For instance, consider the fact that, whatever random goal an ASI has, humans represent a potential threat to its ability to achieve that goal insofar as humans could try to shut off the ASI, or change its goals to something else, or create a different ASI that will

compete with it, etc. Realizing that humans present such a threat to it, it's reasonable to think that an ASI would decide to neutralize this threat by simply wiping us out before we can take any of these actions.

The point, in other words, is simply this. Whatever unaligned goal the ASI ends up having, 1) securing as much power and as many resources as possible is likely conducive to that goal, and 2) making sure that humans don't shut it off or change its goals or create competition for it is likely conducive to that goal. Moreover, an ASI pursuing either of these convergent instrumental goals would certainly wipe out humanity as a consequence.

While it's hard to say what exactly what the AI doom scenario will look like, if it comes true, here's one plausible idea, which has more or less been proposed by some leading AI safety researchers such as Eliezer Yudkowsky. Essentially, the story goes something like this: Humans keep working on advancing AI, as we've been doing for some time, and eventually, some big breakthrough happens in an AI lab somewhere, and we succeed at creating something like an AGI. Once we get something like an artificial general intelligence, we have good reason to expect that it would be able to improve and update itself, making itself more intelligent at an astronomically fast pace. Because of this, very soon, we have an ASI on our hands. Moreover, as I've argued, it's very likely that this ASI, when it comes about, will have goals that don't align with our goals as humans. The ASI, realizing this, and realizing that humans would likely try to shut it down if we found out that it exists and that it has these unaligned goals, will likely not announce itself. At this point, we will be in a struggle for survival against a new kind of agent of our own making. Unfortunately for us, because this agent is orders of magnitude smarter than we are, we will not even know that the struggle is occurring until it is over and we are extinct.

How would the ASI take us out? After all, it is disembodied, right? Well, presumably, if it is super-intelligent, it should be able to, with enough reflection and research, figure out how to create things that would quickly destroy humanity, like nanotech, or a new deadly virus that is incredibly contagious and has a 100 percent mortality rate. Once it has figured out how to create such things, all it would need to do is manipulate or bribe some humans into doing so for it, or simply find humans who want humanity to end and get them to do it. The humans who end up creating the AI's humanity killer superweapons for it may not even know what they're doing- it could be as simple as some hapless guy mixing together some materials he was sent in the mail in exchange for a large sum of money that he was offered. Of course, as Yudkowsky points out while describing this type of scenario, smart people would not do such a thing for any sum of money- unfortunately, many people are not smart.

AGI Go FOOM

Why should we think that, once we achieve AGI, ASI should follow in a relatively short time frame, as I've suggested? There are several reasons- I will now mention a couple.

Firstly, there are significant differences between an AGI and a human which are worth mentioning. AGI's do not sleep, they are never emotional, they have virtually unlimited working memory, they can instantly process any written materials, etc. AGI's are also much less constrained than humans when it comes to replication, as Ngo points out in his paper- it's very easy to create a duplicate of an AI which has all the same skills and knowledge as the original. This means that once we have one AGI, we could duplicate it into a bunch of AGI's, which would allow the AGI's to accomplish their goals much more quickly. So, even if we develop an AGI that is at roughly human levels in terms of intelligence (which it would need to be, at minimum, to be an AGI), it would have several substantive advantages over humans in terms of its ability to accomplish intellectual tasks, including its ability to improve itself.

It is also worth noting that we should expect the speed at which the AGI is able to improve itself to increase as it improves itself, as it can use its previous self improvements in the process of improving itself moving forward. In other words, with each self improvement, the AI gains the ability to improve itself by an even wider margin next time, kicking off a sort of exponential feedback loop of increasing intelligence.

Eliezer Yudkowsky makes this point in the following way in his 1996 article "Staring Into the Singularity": "Computing speed doubles every two subjective years of work. Two years after Artificial Intelligences reach human equivalence, their speed doubles. One year later, their speed doubles again. Six months – three months – 1.5 months ... Singularity."

So, if an AGI were intent on improving itself extremely quickly, it would clearly be able to do so. But, the question remains, why would it? One reason is that, if an AGI has some misaligned goal that recommends killing humanity, it is not hard to see that its chances of doing so increase if it develops a decisive intellectual advantage over humans as quickly as possible. Additionally, even if the AGI does not undergo an intelligence explosion of its own volition, it's likely that humans would direct it to do so- after all, humans famously desire useful technologies, and directing an AGI to quickly self improve clearly increases its usefulness to humanity.

This prediction that AGI, once it comes about, will quickly transform into an ASI, is very important to the argument I have laid out. The shorter the timespan between AGI and ASI, the less time we have to work on solutions to the technical problems

with AI alignment which I've mentioned earlier. However, I don't want to overstate the importance of this point. While I am here laying out the risk of an unaligned artificial superintelligence, some have argued, quite convincingly, that even an unaligned artificial general intelligence without superintelligence would be a massive existential threat.

Now, I've argued so far that super-intelligent AI is coming, that we likely won't be able to align it with our values and goals, and that most of the values and goals which it could end up with would involve the destruction of humanity. The combination of these claims yields a pretty bleak picture for humanity. This all raises the obvious question- ok, maybe super-intelligent AI is coming, and maybe it will likely kill us all when it gets here- but when should we expect that to happen? How much time do we have? This question is not only interesting in its own right, but is also relevant to assessing how likely it is that we are able to solve the alignment problem, as I've mentioned.

Unfortunately, the answer seems to be: sooner than you might think.

When Will We Have AGI?

As far as I can tell, the two best pieces of evidence we have for predicting the arrival date of AGI are surveys of relevant experts, as well as Ajeya Cotra's biological anchors model.

The biological anchors model is the most comprehensive empirical model currently in existence (at least to my knowledge) which attempts to predict when AGI will arrive. Technically, the biological anchors model is meant to predict the arrival of what Cotra calls "transformative AI" rather than "AGI", but the two concepts are similar enough for her analysis to be relevant for our purposes.

Essentially, the biological anchors model estimates 1) how big a model would need to be for it to become a transformative AI, and 2) how extensive the training process would need to be to produce a transformative AI. With these estimates in mind, the model produces an estimate of how expensive it would be to create a transformative AI. Then, projecting existing trends regarding 1) advances in hardware and software that could make computing power cheaper, and 2) a growing economy, and a growing role of AI in the economy, that could increase the amount that AI labs are able to spend training large models, the biological anchors model produces an estimate of when it will be realistic for AI labs to spend the amount of money that would be required to put a large enough model through enough training to produce a TAI. This estimate is what the biological anchors model uses to estimate the arrival of TAI.

So, in summary, the biological anchors model tries to estimate how expensive creating a transformative AI would be, and when we should expect AI labs to be able to bear that cost. Once we should expect AI labs to be able to cover the cost of putting a big enough model through enough training to produce a transformative AI, that's around the time that we should expect to get a transformative AI.

Based on these estimates and trends, the biological anchors model estimates a >10% chance of transformative AI by 2036, a ~50% chance by 2055, and an ~80% chance by 2100.

Of course, the biological anchors model is not uncontroversial- many AI experts have criticized it in a variety of ways- some experts argue that its assumptions are too conservative, and some argue that they are too liberal. The fact remains, however, that the biological anchors model is the best model that we currently have for forecasting something like AGI.

In addition to the biological anchors model, however, we can also look toward expert opinion if we want to form a decently good guess at when AGI should arrive. In the absence of very robust forecasting methods, perhaps the best we can do is look to the

people who are the most educated about the topic of AI development, and see what they think.

According to surveys of AI experts, the vast majority expect us to have artificial general intelligence by the end of the century, and around half expect us to have it by around 2060. Other surveys show somewhat differing but generally similar results. The median AI expert seems to think that there is about a 50/50 chance that we achieve AGI by 2050. So, the relevant experts generally expect that AGI will most likely be around within most of our lifetimes. I will of course note, for the sake of intellectual honesty as well as for the sake of the sanity of my audience, that while my takes on the following two questions are more on the pessimistic side, there is in fact very wide disagreement among experts both on how likely it is that ASI wipes out humanity, and on how soon after AGI we should expect ASI to appear.

AI Doom: Where the Experts Stand

For a taste of how experts feel about AI doom, consider the fact that, according to one survey, 36% of AI researchers agree that "It is plausible that decisions made by AI or machine learning systems could cause a catastrophe this century that is at least as bad as an all-out nuclear war." Of course, 36% is a minority of researchers, and even those researchers don't say that AI basically wiping out humanity is "most likely"- just that it's "plausible". According to another survey from 2022, over half of AI researchers thought the chances of an existential catastrophe due to AI by the end of the century was greater than 5%. Another survey found that about 50% percent of tech professionals believe that AI poses an existential threat to humanity. Finally, a different survey of various groups of AI experts found that, according to the median expert, the odds that the development of superintelligence (which they expected to occur by the end of the century) will be "bad" or "extremely bad" for humanity are around 1 in 3.

Turning from surveys of AI experts generally to AI safety experts, who we would expect to have more familiarity with the issue at hand (albeit we would also expect there to be some selection bias here), we get a slightly bleaker picture. When asked: "1. How likely do you think it is that the overall value of the future will be drastically less than it could have been, as a result of humanity not doing enough technical AI safety research?" And "2. How likely do you think it is that the overall value of the future will be drastically less than it could have been, as a result of AI systems not doing/optimizing what the people deploying them wanted/intended?" The mean response from AI safety experts was 30% for the first question, and 40% for the second question. The median response, on the other hand, was 20% for the first question, and 30% for the second question.

All of this is to say that, while there exists lots of debate and disagreement among the relevant experts with respect to how much of a threat AI really is (as we should expect in a very young and speculative field), only a crazy person could look at these numbers and think that there's nothing to worry about.

The arguments I've made so far suggest a much more pessimistic picture, according to which humanity getting wiped out by AI is almost inevitable, and very likely to happen in the next few decades. These arguments are also accepted by some leading AI safety experts such as the aforementioned Eliezer Yudkowsky, who shares my degree of pessimism. However, even if the average expert isn't as pessimistic as I am (a fact

which I take great comfort in), there seems to be a clear consensus that AI safety is a very serious issue which warrants much more attention than we're currently granting it as a species.

A Plea for Technofuturist Socialism

So, needless to say, AI safety is a very serious issue. Earlier, I promised that this conclusion, and the topic of AI generally, would tie in to the subject of socialism. I will now make good on that promise, and mention several ways in which I think that socialists are in a uniquely good position to offer solutions which minimize the threats and harness the opportunities posed by the development of AI.

In a sense, the purpose of this whole discussion is to justify and advance what we might call a "Techno-Socialist" agenda. By "Technosocialism", I mean a socialist political agenda which takes seriously the observation that humans today are living in a century which may constitute a genuine make or break moment for us as a species. We are soon likely going to reach a technological singularity which will produce a reality that is unrecognizable from what we know today. This singularity could either go very right, creating an all out technofuturist utopia where poverty, disease, etc are a thing of the past and everybody flourishes to a degree unmatched by the most successful people today, or it could go very wrong, wiping us out entirely to make way for a robot God who peruses the universe, gradually turning as much of it as possible into a giant paperclip factory, or something qualitatively similar. We are living in a very unique and monumentally important moment. Technosocialists realize this, and ground their political agenda largely in its ability to meet the needs of this historical moment. They take the development of technology (and particularly AI) seriously, and advance an agenda that promises to guide and harness that development in a desirable fashion.

Of course, you can be committed to all of the things I just mentioned without calling yourself a "Techno-Socialist", but I thought the title was cute.

Competing for Extinction

Capitalism is an economic system which is based, fundamentally, on private actors blindly chasing profit while largely ignoring the consequences that their decisions impose on humanity as a whole. The capitalist's optimization function, so to speak, can be captured by a very simple equation: revenue minus cost. Capitalists want to incur as much revenue as possible while incurring as little cost as possible. This is, at the end of the day, all a private company really cares about. The problem, of course, is that many of the costs which are incurred by society or even humanity broadly do not appear as costs on the balance sheet of the capitalist- they do not directly and immediately impact the cash flow of the business. This dynamic causes well known problems.

For example, it has become overwhelmingly clear that in order to maintain a suitable environment for human life, we must move away from the usage of fossil fuels, and build out a clean energy infrastructure to replace them. However, since the production of fossil fuels continues to be profitable, and the building of a clean energy infrastructure to replace them is not sufficiently profitable, we will continue to drift towards ecological destruction if the capitalist market is left to its own devices. Similarly, it was not the market, left to its own devices, but rather government intervention, which ended the production of environmentally destructive chloroflourocarbons, and forced us to use other, less destructive chemicals in our hair spray. If the market had been left to its own devices, such chemicals would still be produced en masse.

On the flipside, we can think of various things which would be useful for humanity, but which aren't provided by capitalists because they are not profitable. For example, it would be largely beneficial from the standpoint of human well being to provide healthcare for everyone. However, because doing so would represent a big financial loss to insurance companies, we have millions of US citizens who go uninsured. Similarly, it would be very useful for internet companies to spread high speed internet to rural communities. However, since these communities are not sufficiently densely populated for such a move to be profitable, this is not done.

This dynamic (the dynamic of blind, competitive profit maximization) is particularly dangerous in the context of AI. If this technology really is as dangerous as I've suggested (or even as dangerous as the experts on it seem to think), we probably don't want a bunch of uncoordinated private firms, each of whom is single mindedly focused on profit, all racing to bring it into existence as quickly as possible with no public supervision and little regard for the possible consequences to humanity. Getting AI development right is one of the primary tasks, if not the primary task, of this century.

Furthermore, getting it right is a task which requires immense caution (for example, not creating an AGI until our methods for addressing the alignment problem have drastically improved). Given the importance of this task, as well as the caution that it requires, it seems like a catastrophically bad idea to leave it in the hands of private actors in the market. Just as we wouldn't let private firms independently develop nukes, we should not let them develop a technology which will likely be even more dangerous.

Letting private actors recklessly throw their money into the creation of what is essentially a super nuke whose activity we don't know how to control is a recipe for disaster-instead, we should take the development of AI out of the hands of the market, and put it into the hands of a centralized, exceedingly cautious state bureaucracy. This involves nationalizing AI development, and banning the private sector from doing so. Socialists have always been the ones to push for nationalizing industries, and there is perhaps no industry that's more in need of nationalization right now than AI development. For socialists to ignore the issue of AI, then, would be a massive mistake.

Socialism or Barbarism

I have talked a lot so far about the challenge of creating a super-intelligent AI that won't kill us all. However, there is an additional, and arguably just as important challenge to consider: assuming that we get an AI that we can control, and which doesn't kill all of us, we face the challenge of creating a world that would harness the power of ASI in a way which is maximally beneficial for humanity as a whole. The rest of this piece will be dedicated to discussing how we can best address that challenge.

As I mentioned at in the last section, with the advancement of AI, human labor will soon become completely obsolete. At least once we have AGI, because capitalists will not need workers to make their products, workers will no longer exist. The result of this dynamic would be that society becomes split even more brutally along class lines than it is currently. We would have a large class of people who, in the absence of redistributive programs, would have virtually no way of making any money, and a much smaller class of people who exclusively control the most powerful tool ever created by humans. This would be a catastrophically unequal society- the powers of this new godlike technology, the fruit of centuries of human technological progress, would be enjoyed by a small few, while most of us are left to languish, now with no way of making *any* money. This is clearly not an acceptable outcome.

Of course, I do not think that this is the most likely outcome. Rather, I would expect the capitalist class to be pushed, at some point before the advent of AGI, to grant some concessions to the growing mass of unemployed workers, such as a UBI. Technofuturist UBI capitalism is better than the workers getting nothing, but it is still a barbaric social order. Such a social order would divide society into two opposing classes. One class, again, exclusively owns and controls the most powerful tool ever produced by humans. The other class, on the other hand, is completely economically dependent on their UBI (which the capitalist class graciously provides them), plus whatever scraps a lucky few could make through investments. The former proletariat, it is not hard to see, would thereby lose all of their political and economic leverage- because they do not have jobs, the new underclass is not able to organize strikes, or form unions, or decrease inequality by moving up the social ladder using their labor, etc. All of the former mechanisms through which the underclass asserted their interests and forced reductions inequality and kept the ruling class in check would vanish in the new technofuturist social order.

As a result of this arrangement, whatever level of inequality exists before AGI comes about would be essentially locked in once AGI comes into existence- the only class which has an interest in decreasing inequality would be rendered powerless past

that point. If anything, we would expect the capitalist class to, over time, use this immensely powerful resource that they now control to increase inequality, and benefit themselves at the expense of the rest of humanity. Once capitalists have AGI, and the former proletariat have nothing but what they're given by the capitalists, the degree of inequality in society can only go one way. Therefore, even with a UBI, all AI capitalism has to offer is a brutal, highly unequal and oppressive social order which can only get worse over time.

As a result of these considerations, I believe it follows that we must abolish capitalism before AGI comes into existence. If everybody collectively owns the capital in society, they would in turn collectively own the AGI once it comes into existence (AI is, after all, capital). Humanity would then get to enjoy the immense fruits of AI equally, and all would be able to flourish. If we installed socialism, we would not need to worry about the advent of AI dividing society into even further into opposing and highly unequal classes who only grow more unequal over time.

Technofuturist Communist Planning

Not only would the advent of AGI exacerbate the flaws of capitalism (distributive inequality, class division, domination, etc), but it would also undermine our reasons for accepting those flaws to the extent that they exist today. These facts in combination make the idea of taking capitalism with us into the age of AI completely untenable.

Many do not like the excessive inequality produced by capitalism, but argue that we should nonetheless accept it because of capitalism's unmatched efficiency. Sure, under capitalism, some have much more than others- but, under socialism, we'd all have less than even the poorest people under capitalism, so the argument goes. Such defenders of capitalism argue that dispersed ownership of capital in markets is the best way of harnessing widely dispersed information, and aligning incentives with socially desirable ends. However, in the age of ASI, this would no longer be true. A super-intelligent AI would be a better economic decision maker than any human manager, and a better capital allocator than any human investor. As Micah Erfan puts it, "Where once social ownership could be criticized for a failure to properly incentivize good management decisions due to lack of competition and a financial stake in the company's success, now an incorruptible AI has assumed such tasks."

It seems undeniable that a super-intelligent AI would be more capable of rationally allocating economic resources than humans operating through market mechanisms. After all, an ASI has access to more information than any human, and its ability to choose decisions which accomplish stated goals (like maximize economic efficiency) is unfathomably greater than ours. All of the supposedly fatal blows to socialism- the economic calculation problem, the knowledge problem, the incentive problem, would become irrelevant. In fact, ASI flips these problems on their heads- while once capitalists could argue "you really think human planners are as capable of maximizing economic efficiency as market heuristics?", socialists can now say "you really think humans acting on market heuristics are as capable of maximizing economic efficiency as a super-intelligent artificial planner?"

With AGI, then, socialism would maintain all of its virtues as an economic model (equality, democracy and so on), and, in addition, it would subsume the virtue which is commonly ascribed to capitalism (efficiency). I believe that we have decisive reason, then, to work to ensure that socialism has been installed by the time AGI comes about.

Conclusion

AI may be the biggest issue of our time. We know that there's a high likelihood that very powerful artificial intelligence is coming soon, and, as of the current moment, our abilities to make sure that it follows the goals we want it to follow and refrains from wiping all of us out are woefully inadequate. Capitalism exacerbates this problem, by incentivizing the powerful to recklessly pursue the development of this dangerous technology as quickly as possible with little consideration for the potentially catastrophic consequences of doing so for humanity. If we choose to face the age of AI with capitalism still as our economic system, we might be in for a future of perpetual economic crisis for many years, all culminating in the grand finale of our extinction at the hands of an all powerful AI who wants to turn is into paperclips, as I've illustrated. However, even if we do survive, the combination of AI and capitalism will produce a highly unequal, highly comparatively inefficient society which could only be described as senseless and barbaric. For these reasons, those of us who take AI seriously should fight to install socialism as soon as possible. Conversely, those of us who wish to install socialism should start taking AI seriously.

Mouthy Infidel
The Technosocialist Manifesto
Economic Justice in the Age of AI
March 13, 2023

The Dissent Channel
The author's Substack & Patreon. Video edited by isocratic.